Statistical methods for the analysis and interpretation of airborne particle exposure metrics within a time series

Pirani, Monica

Awarding institution:
King's College London

Statistical methods for the analysis and interpretation of airborne particle exposure metrics within a time series framework

by

Monica Pirani

Analytical & Environmental Sciences Division
Faculty of Life Sciences & Medicine
King's College London

A thesis submitted to King's College London
in fulfilment of the requirements for the degree of
Doctor of Philosophy

February 2016

Declaration

I, Monica Pirani, declare that the work presented in this thesis is my own, except where stated otherwise. Furthermore, I confirm that large portions of chapter 3 and 4 have appeared in the following two published papers:

- Pirani, M., Gulliver, J., Fuller, G. W., and Blangiardo, M. (2014), "Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas," *Journal of Exposure Science and Environmental Epidemiology*, 24, 319-327.
- Pirani, M., Best, N., Blangiardo, M., Liverani, S., Atkinson, R. W., and Fuller, G. W. (2015), "Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles," *Environment International*, 79, 56-64.

Monica Pirani

King's College London, February 2016

Abstract

Current urban air pollution is a major environmental risk to population health. Much of the evidence on air pollution and its effects are based on studies focused on a single pollutant, where co-pollutants are treated as confounders or modifying factors. In reality, polluted air exists as a complex mixture of particles, gases and toxic substances and people experience a simultaneous exposure to multiple pollutants and sources.

This thesis is concerned with statistical methods for characterising exposure metrics of airborne particulate matter (PM) and sources within a time series framework. Two original Bayesian modelling approaches are presented, with application to real-world data and inference based on Markov Chain Monte Carlo methods.

A hierarchical modelling approach, which incorporates temporal and spatial statistical structures, was developed for estimating and predicting short-term concentrations of particles from different sources in an urban environment. Taking advantage of a varying coefficient model, this approach modelled the long-range transport of the secondary PM and local primary components, combining observed concentrations from monitoring networks with output from a local-scale dispersion model, while accounting for factors with direct or indirect influence on the particle distribution and formation.

A semiparametric model, based on a Dirichlet process mixture model defined by a stick-breaking construction, was proposed for clustering time points with similar particle and health response profiles. This model used a one-step procedure for dimension reduction and regression, while adjusting for aspects associated with

time variation such as trend and seasonality through smooth functions. It also provided a tool to assess the changes in health effects from various policies to control ambient PM.

These models are flexible and reproducible in different environmental contexts, and were able to capture dependencies in real data and predict temporal and spatio-temporal responses with associated uncertainty.

Acknowledgments

Many people have both helped and supported me, in different ways, to the completion of this research project.

First of all, I would like to express my heartfelt gratitude to my principal supervisor Dr. Gary Fuller for having been my guide throughout the course of this project. Thank you Gary for your permanent support and encouragement, for having guided me through environmental science with great enthusiasm and energy, for being so tremendously generous with your time, and again thank you for your trust, allowing me to conduct this research in my own time and in my own way.

I am also very thankful to my two supervisors Dr. Richard Atkinson and Prof. Nicky Best for the invaluable suggestions, their insightful comments and the helpful feedback regarding my research project.

I owe my special thanks to Dr. Marta Blangiardo for being a constant source of encouragement for me and for being there to keep me going in times of doubt, not only in my research but also in my professional life. Thank you Marta for your mentoring on Bayesian statistics since I started working at Imperial College London as well as along the course of my PhD, and sincerely thank you for the confidence that you had in my potential.

I wish to thank the other co-authors of the papers produced within my research project. In particular, thanks to Dr. Silvia Liverani for the precious comments and insight, and also for reading my dissertation providing helpful suggestions for its improving; and to Dr. John Gulliver for the productive and stimulating collaboration, and for sharing the atmospheric modelled data for London.

I also wish to thank Dr. Ulrich Quass for sharing the data from North Rhine-Westphalia (Germany) and Prof. Stephen Sturzenbaum for the administrative support during my PhD.

I would like to take this opportunity to thank all the friends and colleagues of the Environmental Research Group of King's College London, for providing a very warm and supportive environment for my doctoral research.

I am sincerely grateful to Prof. Sujit Sahu not only for being comprehensive and supportive during my year of work at the University of Southampton, allowing me to take time out for working on the thesis, but also for the invaluable insight into hierarchical modelling. I would also like to thank my dear colleagues Dr. Chigozie Edson Utazi and Dr. Sabyasachi Mukhopadhyay at the S3RI. Thanks Ed and Saby for your friendship, I will miss our laughs, discussions and reflections.

I am forever indebted to my friends in London and in Italy that have supported me throughout these years. It is not possible to mention everyone here but my deepest thoughts are for everyone of you.

My PhD was funded by the Medical Research Council and Public Health England (MRC-PHE) Centre for Environment & Health and I am deeply grateful for its support that made it possible for me to conduct this research project. Moreover, I would like to thank the Traffic Project in London for supporting the publication of one paper.

Last but not least, to Fabrizio, Lara and my beautiful family, impossible to put any acknowledgement in words, but just to say thank you for your love and for always being my pillars of support.

Table of Contents

List of Tables

List of Figures

1 | Introduction

1.1 Motivation

Ambient air pollution exists as a heterogeneous mixture of compounds with a range of physical and chemical properties, consisting of solid and liquid particles, gases, toxic and non-toxic substances, which derives from different (anthropogenic or natural) sources and is the effect of atmospheric transformation and reactions.

Worldwide scientific evidence has identified polluted air as a major environmental risk to health, and therefore a primary regulatory and public health concern. Human and animal toxicological studies (e.g., Stanek et al. 2011a; WHO/Europe 2013) and observational epidemiologic studies (e.g., WHO/Europe 2004; U.S. EPA 2012; WHO/Europe 2013) have largely supported the long- and short-term adverse health effects of air pollution. Recently, evidence has been judged sufficient to classify outdoor air pollution as carcinogenic to humans (IARC 2013; Loomis et al. 2013). An evaluation of the burden due to polluted air, quantified by the the effect of particulate matter (PM) which is a major component of outdoor air pollution, showed a significant reduction in life expectancy of the average population by approximately a year in Europe (WHO/Europe 2006), and it has been estimated that it causes more than 3.7 million of deaths per year worldwide (Anenberg et al. 2010; WHO 2014).

Most of the studies on air pollution exposure and health effects have traditionally considered responses to individual pollutants, and have treated the co-pollutants as modifying or confounding factors.

The reliance on single pollutant or source results is due, in part, to measurement

and source complexities which have limited the development of statistically robust multipollutant models, and in part due to the regulatory strategies of air quality management which have addressed a single pollutant at a time (Dominici et al. 2010).

However, while these single pollutant studies are important in a regulatory context, they do not reflect the real complexity of air pollution exposure that consists of a complex mixture of different compounds from different sources changing in time and in space. Because people are exposed to a mixture, the adverse health effects studied might involve both, individual pollutants and mixtures of contaminants coming from a range of sources, that interact themselves and via complex biological mechanisms. Therefore, estimation of how simultaneous exposure to multiple air pollutants and sources affects the risk of adverse health responses, represents a challenging task for scientific research and air quality management.

To gain better insight into the features of air pollution exposure and its effect, different U.S. and international agencies have declared the assessment of the health effects of pollution mixture a research priority (HEI 2002; National Research Council 2004; WHO/Europe 2007; U.S. EPA 2008). This call to switch from a single pollutant to a multipollutant approach has been embraced by the scientific community (e.g., Mauderly et al. 2010; Dominici et al. 2010; Vedal and Kaufman 2011), with the aim of: (i) accurately characterising the complexity of the air polluted exposure and its impact, (ii) identifying the most harmful pollution emission sources and compounds, and (iii) supporting effective air quality strategies and policies of control.

The analysis of exposure as made up of multiple metrics of pollutants and/or sources, as well as the quantification of the magnitude of their simultaneous effect, represents a challenging aspect of research. Furthermore, the use of epidemiologic results for policy strategies of air pollution control, places heavy weight on statistical methods. Tools for statistical modelling of single or few pollutant(s) are well-developed. However, extending these tools or integrate them with new techniques to model multiple exposure metrics is an active area of research. One of

the challenging issues in modelling air pollution exposure metrics is their intrinsic highly correlated nature, in space and in time, and secondly their interaction with common meteorological and environmental processes. This thesis tackles this challenge.

1.2 Contextual setting

The central subject of this thesis is statistical methods for characterising ambient particle exposure metrics. The focus is in modelling metrics of airborne PM constituents and sources.

The framework of the analysis is constituted by a specific study design, the time series analysis, which considers data collected through time. In environmental epidemiology, time series analysis is used for assessing the short-term exposure health effects of polluted air.

The statistical paradigm adopted in developing the main contributions of this thesis is Bayesian, and both parametric and nonparametric models are explored, presenting in turn statistical structures in which the model has a form that is expressed by a finite number of parameters and in which the model has a form that is expressed by infinite many parameters.

A short description of the physical and chemical characteristics of airborne PM is provided here, along with a picture of the time series framework and the Bayesian paradigm used for modelling, uncertainty assessment and predictions.

1.2.1 *Object*: airborne particulate matter

The term PM is used to describe a complex mix of extremely small airborne particles in solid or liquid form. These particles can contain organic and inorganic substances and vary widely according source, size, shape, solubility, chemical composition, optical properties.

PM has contributions from both *primary* sources emitted directly into the atmosphere, and *secondary* processes formed in the atmosphere by transformation

involving different precursor gases (mainly sulfur dioxide (SO_2), oxides of nitrogen (NO_x), ammonia (NH_3) and non-methane volatile organic compounds, that originate from either local or long-range sources), which produce substances that condense or nucleate into solid or liquid phase becoming PM (McMurry et al. 2004; Kelly and Fussell 2012).

Primary particles and secondary particle precursors originate from a variety of anthropogenic and natural sources. The former includes mainly fuel combustion from both stationary (e.g., power plants, industrial boilers, residential heating and cooking) and mobile sources (e.g., cars, trucks, buses, trains, marine vessels, airplanes). Other sources such as road dust, agricultural emissions, biomass burning and manufacturing processes also contribute. The natural sources of PM emissions include: windblown dust, salt-spray formation in oceanic and sea breaking waves, natural gaseous emissions, ash from volcanic activity, pollens, fungal spores, soil particles and forest fires.

The physical and chemical composition of PM is largely determined by its source, but also depend on location, time of year, and meteorological factors, such as wind speed and direction, temperature, sunlight and relative humidity (Seinfeld and Pandis 2006).

Atmospheric particles have a variety of physical properties that include shape and size. The shape can largely vary: liquid droplets, regular or irregular shaped crystals or aggregates of odd shape. The size presents different orders of magnitude, ranging from few nanometers (nm) to over 100 micrometers (or micron, μm). Because of the irregular shape, a common method of PM measure is represented by the aerodynamic diameter, defined as the diameter of a sphere of unit density (1.0 g cm^{-3}) that has the same settling velocity of the particle under consideration (Finlayson-Pitts and Pitts 2000). Denoting the aerodynamic diameter with D_a, it is computed as:

$$D_a = D_g h \sqrt{\frac{\rho_p}{\rho_0}}$$

where D_g is the geometric diameter, ρ_p is particle density, ρ_0 is the unit particle

density (1.0 g cm^{-3}), and h is the particle shape factor (that is 1 g cm^{-3} in the case of a sphere).

Aerodynamic size is typically used for differentiating PM, because it is partially associated to its generative process, it governs the transport and removal of particles from the air and also it largely determines how it enter in the human respiratory tract. The distribution of particle sizes within an aerosol is called the size distribution. The literature defines the aerosol particle size distributions differently.

A mode classification of particles was originally proposed by Whitby et al. (1972) and Whitby (1978), based on the size distributions and formation mechanisms. According to this classification, in urban environment the distribution of PM tends to be of three-modes: *coarse*, *accumulation*, and *Aiken nuclei* modes (see also U.S. EPA 1996; Baron and Willeke 2001). The coarse mode, larger than 1 μm, regards particles usually mechanically generated (e.g., from wind erosion of crustal materia, construction, and sea spray); (ii) accumulation mode, between 0.1 and 1 μm, includes particles formed by coagulation and condensation of low-volatile gas; and (iii) Aiken nuclei mode, smaller than 0.1 μm, includes particles that are generated by combustion or atmospheric transformation and contributes to the majority of particle number concentrations (PNC). There are also nucleation or ultrafine mode particles (that is, condensation of low vapour-pressure substances formed by high-temperature vaporization or by chemical reactions in the atmosphere to form new particles (WHO/Europe 2000)), which consist of the smallest particles in the size distribution with diameters less than 0.01 μm.

An other way to classify PM according its size is given by the occupational health community that refers to how PM penetrates into the human respiratory tract. This definition is adopted, for example, by the American Conference of Governmental Industrial Hygienists (ACGIH) to define the threshold limit values for occupational exposures (ACGIH 1994) and leads to distinguish the size fraction in: *inhalable*, *thoracic*, and *respirable*. In detail, inhalable are those particles with the potential to enter in the respiratory tract, and includes the nasopharynx

and head airways region; thoracic particles have the potential to penetrate in the tracheobronchial (or conducting airways) region, and the respirable particles have the potential to enter in the gas-exchange (or parenchymal, alveolar, or pulmonary) region (e.g. Brown et al. 2013). In general, particles with diameter larger than 10 μm may penetrate in the nasopharyngeal region, while smaller particles may penetrate deeply into the lung. In particular, particles less than 10 μm can deposit in the tracheobronchial regions and particles less than 2.5 μm can penetrate in the alveolar region (Kim et al. 2015).

Currently, a common approach to size classification (e.g., WHO/Europe 2006 and U.S. EPA `http://epa.gov/ncer/science/pm/`, accessed 15 March 2015), specifies:

- PM_{10}, defined as particles equal to and less than 10 μm in aerodynamic diameter;
- $PM_{2.5}$, also known as fine fraction particles, defined as particles with an aerodynamic diameter of 2.5 μm or less;
- $PM_{10-2.5}$, also known as coarse fraction, defined as particles with an aerodynamic diameter greater than 2.5 μm, but equal to or less than a nominal 10 μm;
- ultrafine particles, defined as particles less than 0.1 μm.

In the thesis, this classification is adopted, as it is largely used in health effect studies as well as for regulatory purposes for what concerns PM_{10} and $PM_{2.5}$. Several studies group particles exceeding $PM_{2.5}$ under the common definition of coarse.

Broadly speaking, coarse particles tend to come mainly from natural sources (including sea spray, pollen grains, mould spores, and plant and insect parts) or by mechanical break-up of larger solid particles from non-combustion sources (including dust from roads, agricultural processes, uncovered soil or mining operations), while fine particles are mainly the result of anthropogenetic combustion processes (including solid and liquid fuel, agricultural field burning, heating

and household cooking, diesel engine combustion, industrial processes); ultrafine particles are formed by nucleation (WHO/Europe 1999; HEI 2002).

The chemical composition of particles varies widely depending on their source and climatology. The major PM compounds are: carbonaceous elements, metals, organic and inorganic compounds, material of biological origin (Putaud et al. 2010; Kelly and Fussell 2012). Secondary soluble inorganic particles are mainly the ion species: sulphate (SO_4^{2-}), nitrate (NO_3^-) and ammonium (NH_4^+), rising from atmospheric reactions. In particular, SO_4^{2-}, NO_3^- and NH_4^+ PM are respectively generated from SO_2 and NO_x and NH_3 precursor gases. The major insoluble components of $PM_{2.5}$ and PM_{10} are organic carbon (OC) and elemental carbon (EC) also called soot or black carbon (BC), that come mainly from combustion processes, notably from road traffic (Harrison and Yin 2000), although some biological sources can contribute to their generation. Metallic components of PM (lead (Pb), cadmium (Cd), vanadium (V), nickel (Ni), zinc (Zn), manganese (Mn) etc.) are generated by metallurgical processes, from impurities in fuel additives and in non-exhaust emissions from mechanical abrasion such as brake- and tyre-wear on vehicles (Kelly and Fussell 2012). Organic compounds are constituted by hundreds of elements, including alkanes, aromatic hydrocarbons, alcohols, organic acids etc., and are emitted from different sources such as diesel engine exhaust, combustion of unleaded gasoline, the effluent from meat cooking operations, and cigarette smoke (Schauer et al. 1996).

The chemical composition varies also according the size of the particles. In general, particles bigger that 2.5 μm consist mainly of insoluble crust-derived mineral, biological material and sea salt, while fine and ultrafine particles are composed mainly from carbonaceous material, metals, secondary particles and organics (HEI 2002).

Recent time series health effect studies have shown that the chemical composition and physical properties of PM play an important role in understanding and discerning the impact of PM on human health, encouraging the hypothesis that

no single component is responsible for the harmful nature of PM (e.g., Bell et al. 2009; Peng et al. 2009; Zanobetti et al. 2009; Atkinson et al. 2010; Bell et al. 2014; Pun et al. 2014), thus supporting a mixture approach to the analysis of environmental exposure metrics.

1.2.2 *Framework*: time series

Time series analysis has the primary objective to develop a mathematical model that provides a plausible description for a set of observations taken sequentially over time (Shumway and Stoffer 2011). In order to provide a statistical setting for the data, the time series can be assumed as a realization from a stochastic (i.e., random) process.

A stochastic process is a collection of random variables, whose members can be identified or indexed by some metrics (Schabenberge and Gotway 2004). Thus, a time series $\{Y_t, t \in \mathcal{T}\}$ is a family of random variables that are ordered in time and are defined at set of time points (Chatfield 2004). $\mathcal{T}$ is called its parameter set. If $\mathcal{T}$ consists of the integers (or a subset), $\mathcal{T} = \mathbb{Z}$, the process is called a discrete time stochastic process; if $\mathcal{T}$ consists of the real numbers (or a subset), $\mathcal{T} = \mathbb{R}$, the process is called continuous time stochastic process. This thesis is concerned with discrete-time sequences.

An alternative useful definition of stochastic process is as generalization of a probability distribution to functions (Rasmussen and Williams 2006). Thus, a probability distribution describes random variables which are scalars or vectors (in case of multivariate distributions), while a stochastic process governs the properties of functions.

In a time series process, the *mean function* and the *autocovariance function* are respectively:

$$\mu_t = E(Y_t) \text{ for all } t \in \mathcal{T} \tag{1.1}$$

$$\gamma(t,j) = \text{Cov}(Y_t, Y_j) = E[(Y_t - \mu_t)(Y_j - \mu_j)] \text{ for all } t \text{ and } j \in \mathcal{T} \tag{1.2}$$

The autocovariance function is also defined as *second moment product.*

After normalisation, the *autocorrelation function* can be obtained, which measures the linear statistical dependence between the members of the time series, that is:

$$\rho(t,j) = \frac{\gamma(t,j)}{\sqrt{\gamma(t,t)\gamma(j,j)}} \tag{1.3}$$

with $\rho(t,j) \in [-1,1]$. Autocorrelation is also sometimes called *lagged correlation* or *serial correlation*, as it refers to the correlation between members of a series of numbers arranged in time.

A sufficient condition for (1.2) and (1.3) to exist is that $\text{Var}(Y_t) = E[(Y_t - \mu_t)] = \sigma_t^2$ to be $< \infty$ for all $\mathcal{T}$ (Cressie and Wikle 2011).

In a time series analysis, one can be interested in measuring the predictability of a time series from another one. Given two time series processes, say $\{Y_t\}$ and $\{Z_t\}$, a *cross-covariance* and *a cross-correlation* can be respectively defined as follows:

$$\begin{aligned} \gamma_{Y,Z}(t,j) &= E[(Y_t - \mu_{Y_t})(Z_j - \mu_{Z_j})] \\ \rho_{Y,Z}(t,j) &= \frac{\gamma_{Y,Z}(t,j)}{\sqrt{\gamma_{Y,Y}(t,t)\gamma_{Z,Z}(j,j)}} \end{aligned} \tag{1.4}$$

Some further characteristics of time series processes that are useful for the analyses developed in the following chapters are now briefly discussed. The section concludes with the description of time series analysis used in environmental epidemiology for studying the short-term health effects of air pollution.

Stationarity

Stationarity is a very important property of stochastic processes. It refers to the stability of the statistical properties of the process through time. There are two widely used definitions of stationarity for time series processes provided here, accordingly to Prado and West (2010).

A time series $\{Y_t\}$ is said to be *strict* (or *strong*) *stationary* if for any lag h and any sequence of times $t_1, t_2, \ldots, t_T$, the joint probability distribution of

$\{Y_1, Y_2, \ldots, Y_T\}$ is identical to that of $\{Y_1 + h, Y_2 + h, \ldots, Y_T + h\}$. In simple terms, this means that the joint distribution of random variables of a strictly stationary stochastic process is time invariant.

A time series is said to be *second order* (or *weak*) *stationary* if, for any lag h and any sequence of times $t_1, t_2, \ldots, t_T$, all the first and second joint moments of $\{Y_1, Y_2, \ldots, Y_T\}$ exist and are equal to the first and second joint moments of $\{Y_1 + h, Y_2 + h, \ldots, Y_T + h\}$. In particular,

$$\begin{aligned} E(Y_t) &= \mu \\ Var(Y_t) &= E[(Y_t - \mu)] = \sigma^2 \\ \text{Cov}(Y_t, Y_j) &= \gamma(j - t) \end{aligned} \tag{1.5}$$

In other words, a second order stationary time series must have three features: constant mean and variance and autocovariance that depend on t and j only through their difference $|j - t|$. Thus, the autocovariance can be written as a function of a particular time lag h, that is,

$$\gamma(h) = E[(Y_t - \mu)(Y_{t+h} - \mu)] \tag{1.6}$$

The necessary and sufficient condition for a function to be an autocovariance function of a stationary time series is to be even and nonnegative definite (Fan and Yao 2003).

The autocorrelation function for a stationary process is:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \tag{1.7}$$

The plot of the autocorrelation function as a function of the lag h produces the so called *correlogram*.

For processes for which the first two moments exist, strong stationarity implies second order stationarity, but not vice versa, as the assumption of finite variance is not assumed in the definition of strong stationarity.

From the definition above, results that both $\gamma(\cdot)$ and $\rho(\cdot)$ are even functions,

that is $\gamma(-h) = \gamma(h)$ and $\rho(-h) = \rho(h)$.

In the situation where two time series are characterised by a stationary behaviour, the cross-covariance and the cross-correlation can be specified as follows:

$$\begin{aligned} \gamma_{Y,Z}(h) &= E[(Y_t - \mu_Y)(Z_{t-h} - \mu_Z)] \\ \rho_{Y,Z}(h) &= \frac{\gamma_{Y,Z}(h)}{\sqrt{\gamma_Y(0)\gamma_Z(0)}} \end{aligned} \tag{1.8}$$

Some simple examples of stationary and non-stationary (that is, the series does not have a constant mean or variance) processes, that have been widely applied and that will be further discussed in this thesis, are now described.

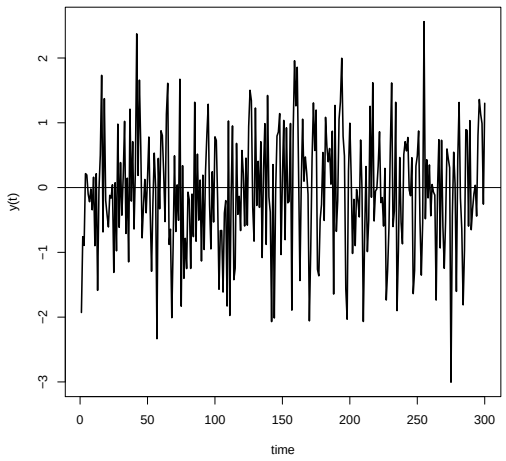
An example of a stationary process is the *white noise process*, that is defined by the following conditions:

$$E(Y_t) = 0, \quad \mathrm{Var}(Y_t) = \sigma^2, \quad \mathrm{Cov}(Y_i, Y_j) = 0 \quad \text{for all } i \neq j$$

These conditions establishes that the expectation is always constant and equal to zero, variance is constant and the variables of the process are uncorrelated for all lags. The white noise process is stationary, but may be not strict stationary. If all of the finite dimensional distributions are Gaussian, the process is called a *Gaussian process* (see also chapter 3). Because uncorrelated Normal random variables are also independent, a Gaussian white noise process is independent and identically distributed (i.i.d.) Normal (0, σ^2) (Fan and Yao 2003). Fig. 1.1 presents a realisation of a Gaussian white noise process, which shows the lack of any predictable pattern over time. Indeed, past values provide no information about the future since the process has "no memory".

An example of non-stationary process is provided by the random walk of order p, which, broadly speaking, describes how an observation directly depends upon one or more previous measurements plus a white noise. In this case, even if the mean is constant, the autocovariance is not independent of time. In particular, assume a random walk process $\{Y_t; t = 0, 1, \dots\}$ of order $p = 1$. The model is

Figure 1.1: Realization of a Gaussian white noise process from i.i.d. $\mathrm{N}(0, 1)$.



defined as:

$$Y_t = Y_{t-1} + \epsilon_t$$

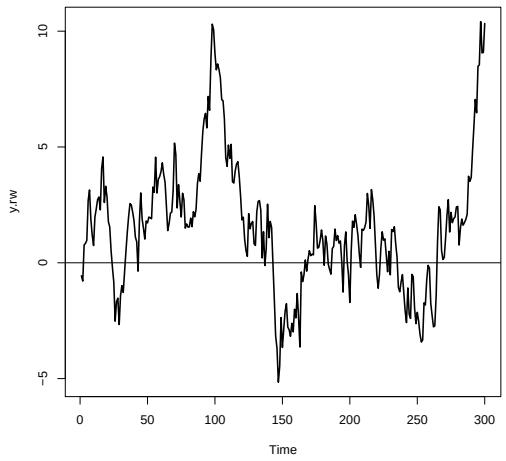
where ϵ_t is white noise process with mean zero and variance σ_ϵ^2. Let Y_0 be fixed. Thus, by recursively substitution starting from $t = 1$, it produces:

$$\begin{aligned} Y_1 &= Y_0 + \epsilon_1 \\ Y_2 &= Y_1 + \epsilon_2 = Y_0 + \epsilon_1 + \epsilon_2 \\ &\vdots \\ Y_t &= Y_0 + \epsilon_1 + \cdots + \epsilon_t \\ &= Y_0 + \sum_{i=1}^{t} \epsilon_i \end{aligned}$$

Hence, $E(Y_t) = Y_0$, which is independent of t, but $\mathrm{Var}(Y_t) = \mathrm{Var}\Big(\sum_{i=1}^{t} \epsilon_i\Big) = \sum_{i=1}^{t} \sigma_\epsilon^2 = \sigma_\epsilon^2 t$ depends on t, thus the random walk process $\{Y_t\}$ is not stationary (see Cressie and Wikle (2011) for more details about features of autoregressive models). Fig. 1.2 shows a realisation of a random walk process, with mean 0 and $\sigma_\epsilon^2 = 1$

In the real-world data analysis, most of the time series are far from stationarity. The natural temporal ordering in the time series creates an internal structure in the data, that shows, commonly, dependence in the observations, such that values

Figure 1.2: Realization of a random walk process.



in the future depend, usually in a stochastic manner, upon observations available at present (Fan and Yao 2003). This means that observations close together in time tend to be correlated (i.e., serially dependent). Time series analysis typically presents challenges, exhibiting patterns behind irregular fluctuations (that is, variations that are short in duration, following not regularity in the occurrence), which include (Chatfield 2004):

Trend, that is the most common time series feature to account for and refers to long-term change in the mean level;

Seasonal variation, which refers to periodic fluctuations which occur periodically within a year;

Cyclic changes, which are recurrent rise and fall that are not of fixed period and are over a period longer than one year.

Most of traditional time series models, however, only work if data are stationary, thus a large body of literature provides statistical methods to deal with dynamic time series data. Chapter 2 illustrates how classical air pollution health effect studies typically handle these time series features.

Time series in environmental epidemiology

In *environmental epidemiology*, time series studies have been extensively used to assess the association between air pollution and short-term health effects (e.g.,

Bell et al. 2004, 2013; Atkinson et al. 2014; and references therein). The London smog episode in December 1952 is probably the most known historical example of time series analysis, where a five-fold increase in death rates was found associated with air pollution episode of four days, bringing the relationship between air pollution and health to the attention of public media, scientific community and government (e.g., Logan 1953; Scott 1953). In the following years, and in particular since the 1990s, time series studies have played an important role in setting standards for acceptable levels of ambient pollution (e.g., Department of Health 1998; WHO/Europe 2004, 2013).

Observational time series studies are termed as *ecological* as they are conducted on communities (such a city) or groups rather than on individuals. This type of study design is used, in fact, to analyse changes in population-averaged acute health outcomes in relationship to short-term variations (typically from zero to six days) in ambient air pollution concentrations.

Time series in air pollution research consist typically of mortality and morbidity data (such as emergency department visits or emergency hospital admissions for cardiovascular and respiratory diseases) collected from administrative databases and metrics of exposure derived from averaged concentrations measured at one or a few air quality background monitoring stations within the study region. Thus, the underling key assumption of this approach is that pollution concentrations are spatially homogenous within the spatial area used for analysis and that the monitor concentration on a given day (or the average of a few monitors) is approximately equal to the true ambient average concentration experienced by the population in study (e.g., Peng and Bell 2010; Bell et al. 2011).

Because time series analyses estimate associations between day-to-day variations in the exposure and day-to-day variations in health adverse outcomes within a specific geographical location, they might be confounded by factors changing on short-time scales (Peng et al. 2006; Peng and Dominici 2008; Bhaskaran et al. 2013). Therefore, typically statistical analyses carefully adjust for *measured factors*, such as weather conditions (e.g., temperature, relative humidity or dew

point temperature) and day of the week, and *unmeasured factors* such as long-term trends which might be attributable to factors like changes in the overall health conditions (due, for example, to improvement in medical practices), sizes and characteristics of the population, seasonal variations in the health outcome and influenza epidemics (Peng et al. 2006; Dominici et al. 2000).

On the other hand, because time series studies focus on variation with time over relatively short periods, many personal characteristics and individual risk factors (such as age, diet, smoking habits etc.) do not change (and also there is no reason to believe that daily pattern of these habits are influenced by air pollution), thus they are unlikely to be potential confounders (e.g., Burnett et al. 2003; Sheppard et al. 2012).

A principal limitation of epidemiologic time series studies is associated with the exposure measurement error, that broadly speaking refers to a situations where observed measurements do not represents exactly the quantity of interest (e.g., Armstrong 1998). This problem could rise at different levels, given the ecological nature of this study design (e.g., Zeger et al. 2000). A discussion of this issue in time series statistical models is presented in chapter 2 (section 2.2.1). The specific problem of the spatial measurement error in exposure modelling, associated with the assumption of spatial homogeneity in air pollutant concentrations, is explored as part of this research and is addressed in chapter 3.

Moreover, a challenging aspect of time series health studies when aimed to quantify the daily exposures to multipollutant metrics, is represented by the correlation between pollutants, which results in collinearity problems. As deeply argued in chapter 2 (section 2.2.2), in presence of multicollinearity traditional regression analysis may be unstable and can achieve results difficult to interpret (MacLehose et al. 2007). In particular, multicollinearity can affects the regression coefficient estimates of the pollutants considered in the analysis and can lead to an increased inaccuracy, as expressed through bias within these regression coefficients, and in increased uncertainty, as expressed by coefficient standard

errors. Chapter 4 proposes a methodological alternative to standard regression methods to deal with high correlated PM components.

1.2.3 *Paradigm*: Bayesian approach

The main research contributions of this thesis are grounded in a Bayesian inferential framework and Markov chain Monte Carlo (MCMC) sampling techniques (e.g., Gilks et al. 1996) are used in exploring posterior and predictive distributions. In this section an introduction to Bayes' method is provided.

Bayes' rule

Consider a sample of observed data $\mathbf{y} = (y_1, \ldots, y_T)'$ from a distributional model $p(\mathbf{y}|\boldsymbol{\theta})$, depending upon a number of unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_P)'$.

In the Bayesian paradigm (e.g., Gelman et al. 2013) conclusions about $\boldsymbol{\theta}$ are made in terms of probability statements, starting with a model providing a joint probability distribution for unknown parameters, $\boldsymbol{\theta}$, and for data, which can be factored as:

$$p(\boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y}) \tag{1.9}$$

and applying Bayes' Rule, the posterior density or probability if discrete, is:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{1.10}$$

These components are as follows:

$p(\mathbf{y}|\boldsymbol{\theta})$ provides the *data description*, also called *sampling distribution* or *measurement model*. When viewed as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{y}$, it is known as a likelihood function, $L(\boldsymbol{\theta}|\mathbf{y})$, as in classical maximum likelihood estimation. Bayesian analysis relies on the likelihood function to draw its inference on $\boldsymbol{\theta}$, and it is given by:

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{t=1}^{T} p(y_i|\boldsymbol{\theta}) \tag{1.11}$$

$p(\boldsymbol{\theta})$ is the *prior distribution* which summarises the knowledge that is available on $\boldsymbol{\theta}$ prior to the see the observed data, such as reflecting beliefs

about dependence structures in the data. Broadly speaking, the prior distributions can be split into two main types, informative prior distributions which contain information that has been obtained from previous analyses/studies/experiments, and noninformative or diffuse priors which aim to express little or no prior knowledge about the parameters.

$p(\mathbf{y})$ is the *marginal distribution*, also called *marginal likelihood* and is a normalising constant, which is needed so that the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$, is a proper distribution (and integrates to unity). Given the case in which $\boldsymbol{\theta}$ is discrete, the marginal probability mass function is obtained by the sum over all possible values of $\boldsymbol{\theta}$, that is:

$$p(\mathbf{y}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}) \tag{1.12}$$

while if $\boldsymbol{\theta}$ is continuous, the marginal probability density function is obtained by integrating $\boldsymbol{\theta}$ out of the joint posterior, that is:

$$p(\mathbf{y}) = \int p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} \tag{1.13}$$

a process referred to as *marginalization*.

$p(\boldsymbol{\theta}|\mathbf{y})$ is the *posterior distribution*, and represents the update of the prior knowledge about the parameters, as summarized in $p(\boldsymbol{\theta})$, given the observed data.

In practice, the denominator in (1.10) is not needed to be computed (it not depend on the parameters, and it only appears as a normalising constant) and Bayes's rule is often written in the unscaled form:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1.14}$$

Predictions

Bayesian inference refers to obtaining the posterior distributions for the parameters of interest and extracting information about these parameters from the posterior. Once the posterior estimates are obtained, it is possible to perform

predictions about new observations. The predictive distribution of a future observation, e.g., $\tilde{y}$, is the conditional distribution of this new observation given the previous observed data $\mathbf{y}$. This can be obtained by integrating the parameters out of the joint posterior of the parameters and new observations:

$$\begin{aligned} p(\tilde{y}|\mathbf{y}) &= \int p(\tilde{y}, \boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &\int p(\tilde{y}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &\int p(\tilde{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \end{aligned} \tag{1.15}$$

Bayesian parametrics and nonparametrics

Within the Bayesian paradigm (as well as in classical statistical inference), is possible to discern between parametric and nonparametric methods. In Hjort et al. (2010) the two Bayesian approaches are described as follows:

- *Bayesian parametrics* involves models with a finite dimensional set of parameters;
- *Bayesian nonparametrics* comprises models characterised by a really large parameter spaces (priors with unknown densities, regression functions etc.) and by the construction of probability measures over these spaces.

Broadly speaking, a parametric model is a parameterized family of distributions, where the number of the parameters does not depend on the sample size, while a nonparametric model is still a parameterized model but the number of parameters may grow as more data are observed (Orbanz and Teh 2010).

There are many Bayesian nonparametric priors based on infinite dimensional families of probability models (e.g., Ghosal 2010; Lijoi and Prünster 2010; Müller and Mitra 2013). The most popular are: (i) the Gaussian process (Rasmussen and Williams 2006), which has gained large popularity within the geostatistical field (Cressie 1993); (ii) the Dirichlet process (DP) (Ferguson 1973; Antoniak 1974) and the Chinese restaurant process (Pitman 1995) and related priors, which are used as priors on latent class models such as those used in clustering and mixed

membership models; (iii) the Beta process (Hjort 1990) and Indian buffet process (Griffiths and Ghahramani 2006) and related priors which are used as priors for latent feature models.

Among these processes, this thesis considers the Gaussian process, which defines a distribution over functions, in chapter 3 and the DP, which defines a distribution over distributions, in chapter 4. The background material on these stochastic processes is provided in the related chapters.

1.3 Aims

Given the specific research window described in the previous section, the goals of this thesis consist in proposing modelling approaches that can provide an answer to some issues associated with standard statistical methods in air pollution and health time series studies, and specifically:

- to improve air particle exposure modelling for short-term health effect studies using geostatistical approaches to combine information on local and regional scale pollutants;
- to increase our understanding of how different mixtures of airborne particles can affect community health;
- to provide insight into the development of statistical methods for epidemiologic studies aimed at exploring health risks associated with exposure to multiple pollutants and sources.

These aims are pursued drawing together statistical and environmental concepts in defining modelling approaches focused on outdoor pollution within a urban environment.

1.4 Main contributions

The major contributions of this thesis are two original studies developed within the common framework of Bayesian analysis. They can be summarised as:

- *A spatio-temporal hierarchical approach for modelling and predict short-term exposure concentrations of PM in an urban environment, with application to multiple sources of PM_{10} in London.* PM_{10} shows generally homogeneous spatial distribution over cities, however part of its components, such as those from local traffic, are likely to be unevenly distributed. In this study, five hierarchical models were developed, accounting for both the spatial and temporal variability of concentrations. Two main source contributions to ambient measurements were considered: (i) the long-range transport of the secondary fraction of particles, which temporal variability was described by a latent variable derived from rural concentrations; and (ii) the local primary component of particles captured by the output of the dispersion model ADMS-Urban, which site-specific effect was described by a Bayesian kriging. The effect of a set of covariates was also considered, including type of site, daily temperature to describe the seasonal changes in chemical processes affecting local PM_{10} concentrations which are not considered in local-scale dispersion models and day of the week to account for time-varying emission rates not available in emissions inventories. The statistical models were constructed using regression techniques, characterised by space- and time-varying coefficients and assessed using cross-validation procedure. The results indicated that concentration estimates in urban areas benefit from enhancing the city-scale particle component and the long-range transport component with covariates that account for a residual spatio-temporal variation in PM.

- *A semiparametric model for clustering time points with similar particle and health response profiles, with application to different metrics of PM in London.* This study explicitly aimed to evaluate the effects deriving by the exposure to airborne particle mixtures within a classical time series design. Methodologically, the study was based on the DP mixture models, defined by a stick-breaking construction, that links nonparametrically the response

to particle data through cluster membership, while adjusting for seasonality and trend components. The applicability of the statistical model was evaluated using daily time series of a range of particle metrics and respiratory mortality counts for London for the years 2002-2005. The results showed a higher risk of mortality in days characterised by high levels of PM, especially secondary particle concentrations, including inorganic anions as sulphate and nitrate. The model also allowed the prediction of the mortality response under a new scenario of exposure as London experienced in the year 2012. This was performed to assess health impact changes as a result of policies that affected many parts of the particles mix in London during the recent years. The comparison of the posterior predictive distributions of mortality under the exposure scenario in 2012 vs 2005 showed a predicted annual average decrease in respiratory mortality associated with the decrease in air particle concentrations. This feature of the model, provided a new tool to assess the changes in health effects from various policies to control the ambient PM mixtures.

1.5 Organization

The remainder of this thesis has been structured as follows.

Chapter 2 presents a methodological overview of well-established and upcoming statistical methods used in air pollution time series studies. It describes the classical regression-based approaches, thus points out the elements of uncertainty and the methodological issues which arise in dealing with multiple components and/or sources of exposure in time series analysis, mainly represented by correlation among pollutants and exposure measurement error. Thus describes the current methods used in literature for characterising exposure metrics, along with their strengths and limitations.

Chapter 3 presents the spatio-temporal modelling approach for short-term exposure estimates and predictions in urban area. It firstly describes the related

literature, as well as the environmental and statistical perspectives adopted in developing the study, providing also some necessary preliminaries on spatial processes. Then, it proceeds with the description of data from London and the modelling approach adopted. It presents the findings and the aspects related to model comparison, parameters interpretation, and quality assessment of predictions. Finally, it carries out a discussion on the exposure assessment approach proposed.

Chapter 4 presents the Bayesian semiparametric model for clustering and regression, characterised by a joint model for health response and particle metrics, within a classical time series framework. It describes the related works in literature, then presents the data set used for London 2002-2005 and 2012, the model developed and the results. The methodology is further compared to a standard regression-based model where particles were grouped using a distance-based clustering algorithm such as K-means. Finally a discussion of the results is provided, with focus on the advantages of the model proposed over traditional clustering techniques.

Chapter 5 summarises the results achieved in the previous chapters, presents the conclusions and proposes several additional future work directions.

2 | Overview of statistical methods used in air pollution and health time series studies

The goal of this chapter is to present a methodological overview of the approaches commonly used in air pollution health effect studies, within a time series design, with specific emphasis on the methods used for characterising exposure metrics of pollutants and their sources.

Section 2.1 describes the classical time series regression-based models and the associated topic of the confounding adjustment, mainly performed using smooth functions. Subsequently section 2.2 presents the major statistical issues and causes of uncertainty in this study design. Section 2.3 provides a broad panorama of the methods used for characterising air pollution exposure metrics. The chapter concludes with a brief discussion on the described approaches in section 2.4.

2.1 Generalised linear and additive models

The classical statistical approach for estimating short-term health risks associated to air pollution exposure is based on *generalized linear models* (GLMs), which extend linear regression to many types of response variables (McCullagh and Nelder 1989) or *generalized additive models* (GAMs) which represent an extension of GLMs, allowing the identification and characterisation of nonlinear regression effects, while maintaining additivity (Hastie and Tibshirani 1990). Unlike classical linear models, which presuppose a Gaussian (or Normal) distribution for the response, in a GLM or a GAM the distribution of the response may be

any member of the exponential family distributions. In detail, let a single random variable Y (which may be either continuous or discrete), whose probability distribution depend on a single parameter θ. The distribution belongs to the exponential family if its probability density function (or probability mass function) can be written as:

$$p(y, \theta) = h(y)\exp\{\theta' t(y) - a(\theta)\} \tag{2.1}$$

where, θ is the *canonical* or natural parameter of distribution; $t(y)$ is the sufficient statistic (it is called sufficient because the likelihood for θ only depends on y through $t(y)$); $h(y)$ is the underlying measure (counting measure or Lebesgue measure); and $a(\theta)$ is the log normalizer ($a(\cdot)$ is referred as the the log-partition function) that ensures that the density integrate to 1: $a(\theta) = \log \int h(y)\exp\{\theta' t(x)\}dy$. Essentially most of the distributions referred to in this thesis are from the exponential family (including, Normal, Poisson, Gamma, Multinomial, Dirichlet).

In its general form, a GLM involves: (i) a response data vector $\mathbf{y}$; (ii) the predictors (or covariates) $\mathbf{x}$ and coefficients $\boldsymbol{\beta}$, forming a linear predictor $\eta = \mathbf{x}\boldsymbol{\beta}$; (iii) a random component specifying the distribution of the response variable, with mean $E(\mathbf{y}|\mathbf{x}) = \mu$; and (iv) a monotone link function $g(\cdot)$ that relates the mean of the response variable with the linear predictors: $\mu = g^{-1}(\eta) = g^{-1}(\mathbf{x}\boldsymbol{\beta})$[1].

A GLM used in time series studies of air pollution and health, is built on a set of data recorded over time and owns the features specified in the Introduction of this thesis. They include an outcome, y_t, for $t = 1, \ldots, T$, which typically consists of daily morbidity or mortality counts; daily air pollution concentrations, $x_{t1}, \ldots, x_{tP}$, for $j = 1, \ldots, P$; and additional time-varying covariates, $u_{t1}, \ldots, u_{tL}$, for $l = 1, \ldots, L$, to control for the non-linear effects of confounding factors. Typically these regression-based models adjust for time-varying confounding factors including smooth functions of time, weather variables (i.e., temperature, humid-

[1]It is common practice to write GLMs using the inverse function of g. The appropriate choices for g depend on the nature of the response variable.

ity) and day of the week. Because the outcome in these studies is constituted by (mortality or morbidity) counts, it is a common choice to assume a Poisson distribution, and the logarithm is taken as link function $g(\cdot)$. The model specification has the form:

$$\begin{aligned} y_t &\sim \text{Poisson}(\mu_t) \\ \log \mu_t &= \beta_0 + \sum_{j=1}^{P} \beta_j x_{tj} + \sum_{l=1}^{L} f_l(u_{tl}, d_l) \end{aligned} \tag{2.2}$$

where the functions $f_l(\cdot, d_l)$ denote smooth functions of covariates and d_l are parameters controlling the smoothness of their respective functions (e.g., Dominici et al. 2004). GLMs commonly define the smooth functions to be regression splines, such as natural cubic splines or B-splines with a pre-specified number of knots at known locations. In (2.2) β_j represents the estimated change in the logarithm of the population average morbidity or mortality count per unit of change in the pollutant j (while controlling for the effects of other covariates), which might be specified at different lag time in comparison to the event in study. This β_j parameter is generally expressed as the percentage of increase in the health effect for every 10 units increase in the exposure metric.

A GAM is a GLM in which the part of the linear predictors, $\sum_{j=1}^{P} \beta_j x_{tj}$, is specified in terms of a sum of smooth functions of the underlying predictors. The exact parametric form of these functions is unknown, as is the degree of smoothness appropriate for each of them. The additive model has the form:

$$\log \mu_t = \beta_0 + \sum_{j=1}^{P} s_j x_{tj} + \sum_{l=1}^{L} f_l(u_{tl}, d_l) \tag{2.3}$$

where s_j is a nonparametric function. These models are called *additive* because require the calculation of a separate s_j for each predictor, and then add together all of their contributions (James et al. 2013).

GLMs and GAMs have become a standard tool for time series analysis explor-

ing the effect of air pollution on population health, and the scientific evidence produced across the world in environmental epidemiology is largely based on these regression techniques. Examples of application of GLMs or GAMs within a single city or community are extremely numerous (e.g., Schwartz and Marcus 1990; Touloumi et al. 1996; Tenías et al. 1998; Atkinson et al. 2010), as well as examples across cities (e.g., Biggeri et al. 2004; Pun et al. 2014). Remarkable examples are the large multi-city studies such as the US *National Morbidity, Mortality and Air Pollution Study* (NMMAPS) (e.g., Samet et al. 2000a,b; Huang et al. 2005), the European *Air Pollution and Health: An European Approach* (APHEA) (e.g., Samoli et al. 2009), the Latin American *Estudio de Salud y Contaminación del Aire en Latinoamérica* (ESCALA) (Romieu et al. 2012).

From a computational point of view, traditionally GLMs are fitted using iteratively reweighted least squares (McCullagh and Nelder 1989), while GAMs rely on the backfitting algorithm (Hastie et al. 2009) in addition to iteratively reweighted least squares.

2.1.1 Smoothing functions

In regression-based techniques applied to air pollution time series health studies, smoothing functions play a central role. The previous section showed that air pollution, meteorological and unmeasured factors can be entered in these models with non-linear behaviour. In particular, GAMs and GLMs deal with the need of adequately control for the potential non-linear confounding effect of time-varying measured and unmeasured confounding factors. Here the trade-off is primarily in removing the confounding effects but retaining as much as possible un-confounded shorter-term fluctuations, that can be associated with short-term fluctuations in the pollutant exposure.

A variety of different approaches have been used in literature. In the earlier time series studies, simple functions of time and weather variables were used, including time-stratified indicators and polynomial of fairly low degree (e.g., Bhaskaran et al. 2013), while cyclical sequences have been mainly exemplified by Fourier

terms, that is a finite sum of pairs of sine and cosine terms. Recently, the most common adjustment methods consists of smooth functions. GLMs commonly define parametric smooth functions to be polynomial regression or regression splines, such as B-splines and natural cubic splines, with a pre-specified number of knots at known locations; while GAMs include nonparametric smoother such as smoothing splines (that place knots at every data point and are sometimes referred to as full rank smoothers because the size of the spline basis is equal to the number of observations), local polynomial regression smoothers (LOESS) or penalized splines (e.g. Peng and Dominici 2008). In this context, the main advantage of nonparametric over parametric models is their flexibility, as in the parametric framework the shape of the functional relationship between response and covariates is determined by the model, whereas in the nonparametric framework the shape is determined by the data. The performance of these different representations of the smooth functions in air pollution time series studies has been object of discussion in literature (e.g., Peng et al. 2006; Touloumi et al. 2006).

In the following section, a brief description of the most commonly used smooth functions is provided. For descriptive purpose, the theme is set out recalling the classical smoothing problem:

$$y_t = f(x_t) + \epsilon_t \tag{2.4}$$

where (y_t, x_t) is the t-th observation from a response variable y and a covariate x, $f(\cdot)$ is a smooth function, and ϵ_t are i.i.d. random errors. The smooth function could be modeled in different ways, using for example polynomials, B-splines, truncated polynomials etc.

Polynomial basis

The *polynomial* is a simple way in which curves can be represented and it is obtained by raising each of the original predictors to a power. Let m be the order

of the polynomial (that is, the number of coefficients defining the polynomial), and let q be the degree of the polynomial (that is, the highest power defining the polynomial, where $q = m-1$). A polynomial basis has the simple form:

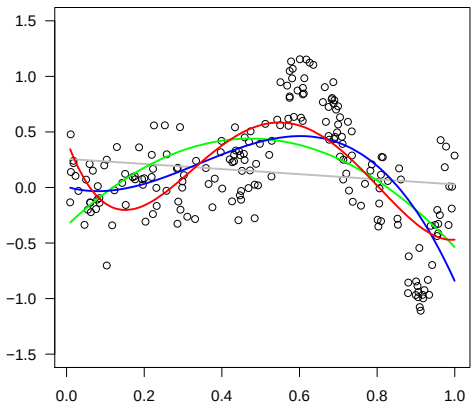
$$f(x) = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3 + \cdots + \gamma_q x^q \tag{2.5}$$

where the coefficients $\gamma_0, \ldots, \gamma_q$ are rational numbers. It becomes the simple polynomial regression model:

$$y_t = \gamma_0 + \gamma_1 x_t + \gamma_2 x_t^2 + \gamma_3 x_t^3 + \cdots + \gamma_q x_t^q + \epsilon_t \tag{2.6}$$

Figure 2.1 provides an example of polynomial regression for some simulated data from the model $y_t = \sin^3(2\pi x_t^3) + \epsilon_t$. The input covariate x is generated by choosing values spaced uniformly in the range [0, 1].

Figure 2.1: Polynomial regression of simulated data from the model $y_t = \sin^3(2\pi x_t^3) + \epsilon_t$; grey is a linear regression, green is a polynomial of degree 2, blue is a polynomial of degree 3 and red is a polynomial of degree 4.



Polynomial curve fitting is a model that is linear in the parameters γ (so polynomial regression is a linear model) even though it is a nonlinear function of the input variable (i.e., predictor/covariate) (e.g., Ruppert et al. 2003; Bishop 2006).

Polynomials are flexible functions but can introduce undesirable side effects. Because polynomial basis functions are global functions of the input variable, each observation affects the entire curve (Bishop 2006). This might introduce

bias, and it also results in extremely high variance near the edges of the range of the variable. Moreover, a polynomial of high degree can be hard to interpret. A solution to this problem is represented by dividing the input space in intervals and fit a different polynomial in each interval, obtaining *spline functions*.

Spline functions

Splines are piecewise polynomials joined together to make a single smooth curve. They are used to approximating nonlinear functions and are constructed by dividing the domain of the variable, say $[a, b]$, into contiguous intervals, then fitting separate polynomials within each range. The polynomials are joined together at the interval endpoints that are called *knots*,

$$a = \xi_0 < \xi_1 < \cdots < \xi_H < \xi_{H+1} = b$$

where the H knots $\xi_1, \ldots, \xi_H$, for $h = 1, \ldots, H$, are called inner knots. A spline of degree q is a function that has the following properties (Turner 2000):

- f is a piecewise polynomial such that, on each interval $[\xi_{h-1}, \xi_h]$, is a polynomial of degree q;
- f is $q-1$ times continuous differentiable at the knots.

In general, a spline of degree 0 is a step function with steps located at the knots. A spline of degree 1 is a piecewise linear function where the straight line segment connects at the knots. A spline of degree 2 is a piecewise quadratic curve with continuous first derivative at the knots. A spline of degree 3 is a piecewise cubic function that has continuous first and second derivatives at the knots.

By requiring continuous derivatives, it ensures that the resulting function is smooth. The smoothness can regulated by increasing/decreasing the degree of the spline and/or the number of knots. This is a critical point, that requires balance between the risks of under-smooth (that is, few knots/low degree which might result in class of functions too restrictive) or over-smooth (that is, many knots/high degree which might produce overfitting). In time series health studies,

under-smoothing can lead to a residual confounding effects and over-smoothing can attenuate the true pollution effect.

There are many ways to parameterize a spline. Usually, they are constructed using spline *basis functions* (e.g., Ruppert et al. 2003; James et al. 2013). Consider (2.4), where for the function $f(\cdot)$ can be given an expression so that it can be written as a linear regression model. This is done by using a family of functions or transformations that can be applied to a predictor x. This means that $f(x_t)$ is built up in basic components, called the *basis functions* $B_r(x_t)$ (of degree q), such that:

$$f(x_t) = \sum_{r=1}^{J} \gamma_r B_r^q(x_t) \tag{2.7}$$

where γ_r's are unknown parameters to be estimated and $B_r(\cdot)$ is a the TxJ design matrix consisting of the basis functions evaluated at specified observations. These functions can be generated in different ways. Some of them as briefly showed below.

Truncated power basis

The truncated power functions (e.g., Ruppert et al. 2003) are defined as:

$$(x_t - \xi_h)_+^q = (x_t - \xi_h)^q I_{x_t > \xi_h}(x_t),\ h = 1, \ldots, H \tag{2.8}$$

where $I_{x_t > \xi_h}$ is an indicator function and the symbol $+$ indicates that the function takes the following values:

$$(x_t - \xi_h)_+^q = \begin{cases} 0, & \text{if } x_t \leq \xi_h \\ (x_t - \xi_h)^q, & \text{if } x_t > \xi_h \end{cases}$$

This involves f being modelled as a function of the form:

$$f(x) = \delta_0 + \delta_1 x + \ldots, + \delta_q x^q + \sum_{h=1}^{H} \gamma_h (x - \xi_h)_+^q \tag{2.9}$$

The truncated power basis is conceptually simple, however it is not too attractive numerically, because the powers of large numbers can lead to severe rounding

problems (Hastie et al. 2009) and also the fitting process can be numerically problematic due to the fact that the truncated power bases are correlated (that is, they are far from orthogonal (Ruppert et al. 2003).

B-spline basis functions

The B-spline basis functions derive from truncated power functions, but in comparison to them, they have more stable numerical properties (de Boor 1978). They solve the issue associated with truncated polynomials as the basis functions are no longer collinear, leading to a more stable numerical fit. In extreme synthesis, a B-spline function extends truncated power function by adding q knots in the interval $[a, b]$ of the covariate in a non-decreasing sequence. de Boor (1978) shows that B-spline of order $m > 1$ are recursively computed from the B-splines of lower order using the recurrence relation:

$$B_j^m(x) = \frac{x - \xi_j}{\xi_{j+m-1} - \xi_j} B_j^{m-1} + \frac{\xi_{j+m} - x}{\xi_{j+m} - \xi_{j+1}} B_{j+1}^{m-1}(x) \tag{2.10}$$

and

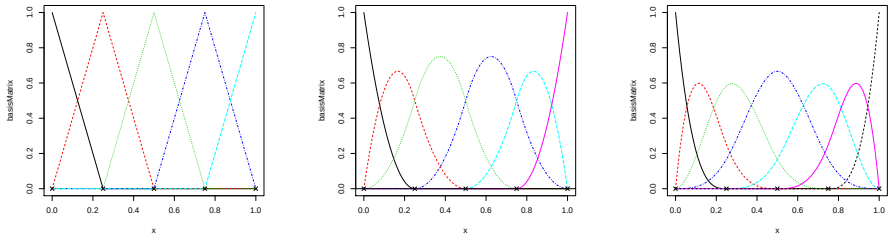
$$B_j^1(x) = \begin{cases} 1 & \text{if } \xi_j \leq x < \xi_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

where B_j^m is the jth B-spline basis function of order m. B-spline of order m has $J = m + H$ basis functions in their formulations, i.e., the number of columns in the design matrix is equal to the B-spline order m plus the number of interior knots H.

Figure 2.2 gives examples for B-spline bases of different orders m, showing the relationship between a basis function and its knot sequence. Thus, since the basis functions are based on knot differences, the shape of the basis functions is only dependent on the knot spacing and not specific knot values.

Once an appropriate set of basis function has been constructed, a spline function estimate of the regression function, f, can be created.

Figure 2.2: B-spline basis function with 3 internal knots at (0.25 0.50 0.75), respectively of (a): order 2, (b) order 3, and (c) order 4.



(a) (b) (c)

Natural cubic spline

Natural cubic splines (also called *restricted cubic splines*) are continuous in the second derivative, at and between knots points (as usual for the cubic spline), and additionally are linear in their tails beyond the boundary knots, i.e., the second derivatives are zero (Turner 2000). Therefore, the function is continuous with straight line outside the interval $[a, b]$ of x, while maintaining its smoothness. Natural splines can be represented by different bases. Among these, the use of B-spline basis matrix is popular.

Smoothing splines and penalised splines

Smoothing splines and penalised splines (Eilers and Marx 1996; Marx and Eilers 1998; Ruppert et al. 2003), are commonly used in GAMs for air pollution health effect studies, where the functional dependence between response and covariates is exposed without imposing any parametric assumption about this dependence.

Spline models using smoothing splines consist in finding the function $f(x)$ as a solution to the following minimisation problem:

$$\min_{f(\cdot)} \left\{ \sum_{t=1}^{T} (y_t - f(x_t))^2 + \lambda \int_a^b (f''(x))^2 dx \right\} \tag{2.11}$$

where λ is the smoothing parameter. This structure consists of two terms. First, the deviation of the fitted function from the observed values should be minimised (this gives a goodness of fit). Second, complex functions are penalised by the second term in (2.11) that is measured by the integrated squared derivative (for

a linear function, this term will be zero). The solution to this minimisation problem turns out to be a natural cubic spline, with knots placed at each data point (Wahba 1990). By changing the value of λ produces changes the smoothness of the estimated function. Smoothing splines rely heavily on the penalty term, so as to reduce the computational cost involved in placing knots at each data point.

Penalised splines can be viewed as a generalisation of smoothing splines, with two main differences. A generous dimension of the number of knots is specified to achieve the desired flexibility, while penalise excess curvature using the penalty term, λ, that is not longer in term of a second derivative, but a second divided difference (Eilers and Marx 1996). In the penalised spline approach also, the smoothing parameter is chosen carefully, commonly by cross-validation. If λ is too high, it can lead to a oversmoothing of the data, while if λ is too low, it can result in an undersmoothing of the data. Thus, because λ controls the amount of smoothing, the value of K is no longer crucial in penalised splines (Ruppert 2002).

Penalised splines have been implemented in a number of forms, for example using B-spline basis (Eilers and Marx 1996) or truncated power-function basis (Ruppert and Carroll 2000), or again radial basis functions (Crainiceanu et al. 2005).

2.2 Uncertainties and issues

Even widely used, classical methods such as GLMs and GAMs present several well known sources of uncertainties, mainly related to:

- the confounding factors considered in the model;
- the degree of adjustment adopted for these factors (e.g., choice of degrees of freedom and knot locations for spline functions);
- the lag structure for the air pollution variables, as the distribution of the effect can be specified at different times after the exposure, commonly

between 0 and six days (on this issue, there is a body of literature focused on considering a distributed lag models, such as Zanobetti et al. (2000, 2002) and Gasparrini et al. (2010)).

In addition, two methodological issues plague the literature on air pollution and health time series design: the potential sources of exposure measurement error and the correlation among pollutants, when including more then one or few pollutants in analysis. These two issues have been considered in this thesis, thus they are discussed more deeply in the two following sections.

2.2.1 Exposure measurement error

The error in the exposure assessment is the difference between the observed or measured exposure from the true exposure. It represents a critical issue in air pollution epidemiology (Armstrong 1998; Zeger et al. 2000; Dominici et al. 2000; Sarnat et al. 2007).

Measurement error of exposure variables (pollutants, potential confounders, or potential effect modifiers) is defined as differential when exposure measurement error is associated with the health outcome, and non-differential when it does not depend on the outcome. The differential exposure measurement error can cause bias in an effect estimate towards or away from the null, while the non-differential typically results in bias towards the null (Armstrong 1998).

In time series health research, the non-differential measurement error has been the focus of discussions, and studies have suggested that the impact of exposure measurement error differs depending upon the type of error introduced. In particular, two possible errors have been identified: the classical or Berkson type, which represent the extremes of a continuum, as most exposure measurement errors combine elements of both (Zeger et al. 2000; Bateson et al. 2007; Goldman et al. 2011; Sheppard et al. 2012).

Let x_t be the true predictors subjected to measurement error, (i.e., the true ambient concentrations) on day t and w_t the observed proxies for x_t (i.e., the

measured ambient levels) on day t. Classical error occurs when true exposures are measured with additive error and the average of many replicate measurements, conditional on the true value, equals the true exposure (Carroll et al. 2006; Bateson and Wright 2010). Specifically, it occurs when it is assumed that measurements, w_t, vary randomly about true concentrations, x_t. This can be considered the case for instrument error associated with ambient monitors: instrument error is independent of the true ambient level, such that $E[w_t|x_t] = x_t$ (Zeger et al. 2000). In this case, the measurement error, $w_t - x_t$, is uncorrelated with the true value x_t. Because the variation in the measurements is expected to be greater than the variation in the true values, classical measurement error tends to bias the true effect toward the null and the effect attenuation will depend on the error variance of the observed exposure relative to the variance of the true exposure (Armstrong 1998; Zeger et al. 2000; Goldman et al. 2011).

Different from the classical error model, Berkson error occurs when part of the true exposure is measured. In this case the average of individuals' true exposures, conditional on the assigned measurement, equals the assigned measurement (Bateson and Wright 2010). Specifically, in a Berkson error model the true ambient, x_t, varies randomly about the measurement, w_t. This might be the case, for example, in which w_t is the spatially averaged ambient level of a pollutant without major indoor sources, and x_t is the personal exposures that match the ambient level when averaged over large populations (Zeger et al. 2000). In this case, measurement is independent of the measured population average over the study area: that is, $E[x_t|w_t] = w_t$. Berkson measurement error will not bias effect estimates but will tend to increase the standard error in the estimates (Zeger et al. 2000; Bateson et al. 2007; Goldman et al. 2011).

There are many sources of measurement error in the analysis of air pollution and health data. Zeger et al. (2000) described a conceptual framework for exposure measurement error in regression-based time series studies, and identify three components of measurement error: (i) the difference between individual exposures and average personal exposure, (ii) the difference between average personal

exposure and ambient levels, and (iii) the difference between measured and true ambient concentrations. The authors observed that the first and the third differences are likely to behave like Berkson error and are unlikely to induce bias, however the variance of the regression coefficient tends to be increased. The second type of measurement error could be, instead, a substantial source of bias.

Recent contributions in literature (e.g., Sarnat et al. 2010; Peng and Bell 2010; Bell et al. 2011) have focused on the third term of the measurement error decomposition of Zeger et al. (2000), considering the aspect of the spatial misalignment error, that is the case of observations collected at different spatial locations. In ecological time series studies this may occur because pollutant data are typically measured at points for monitor stations and health outcomes are often aggregated over the given area. In the traditional approach, the exposure estimates are typically obtained as the spatially averaged ambient pollutant levels, and these aggregated pollutant data are assumed to be representative of the exposure experienced by the study population, therefore compared with aggregated health data. However, if this assumption of spatial homogeneity of the pollutants (that is needed to proceed to this aggregation) does not hold, poor estimates of the ambient averaged concentrations are likely occur, along with poor estimates of the associated health risk (Shaddick et al. 2013).

Sarnat et al. (2010) analysed the measurement error in relationship to spatio-temporal variability in ambient air pollution concentrations measured at different monitoring sites (i.e., central urban and rural sites) in the area of Atlanta (US), comparing the health risk estimates by monitor location. The authors considered pollutants like $PM_{2.5}$ and secondary pollutants, such as O_3, that are reasonably spatially homogeneous, in that their concentrations, as well as the temporal fluctuations in their concentrations, are relatively consistent over a study region, and other pollutants, including those emitted by motor vehicles, such as CO and NO_2, that are likely to show spatio-temporal heterogeneity. They found similar health effect estimates for spatially homogenous pollutants and discrepancy for spatially

heterogeneous pollutants in urban versus rural monitors. Thus, study's findings supported the use of pollutant concentrations from urban central sites to assess population exposures within geographically dispersed cities.

Peng and Bell (2010) described a methodology for addressing the spatial misalignment error, using regression calibration (see also Bateson and Wright 2010) and two-stage Bayesian approach. Analysing the risk of cardiovascular hospitalization associated with exposure to chemical components of $PM_{2.5}$ in 20 urban counties in the US, Peng and coworkers found that the monitor average is good proxy for true value with good monitor coverage and/or low spatial heterogeneity. But, spatial misalignment adjustments are useful when: (i) pollutants are spatially heterogeneous, such as sodium ion (Na^+), silicon (Si), and EC, and (ii) the monitor coverage is poor within the area of interest, but monitors exist outside the area in study so that information about spatial variability of a pollutant can be "borrowed" from outside area by fitting a spatio-temporal model to available data.

Bell et al. (2011) investigated the spatial relationship of seven chemical constituent concentrations of $PM_{2.5}$, for the period 1999-2007, for 480 monitors in the US. They found that spatial heterogeneity was present for all constituents, yet lower for NH_4^+, SO_4^{2-}, and NO_3^-. Lower correlations were associated with higher distance between monitors, especially for NO_3^- and SO_4^{2-}, and with lower long-term levels, especially for SO_4^{2-} and Na^+. Analysis of collocated monitors revealed measurement error for all constituents, especially for EC and Na^+. Thus, the authors concluded that exposure misclassification may be introduced into epidemiologic studies of $PM_{2.5}$ constituents due to spatial variability, and is affected by constituent type and level.

Goldman et al. (2011) considered the combined errors in time series studies of: (i) instrument precision error, and (ii) error due to spatial variability, and assessed the effect of error type (classical and Berkson) on the impacts of measurement error on epidemiologic results from a study of 12 air pollutants and emergency visits in Atlanta (US). The authors found that the health risk estim-

ates of exposure to ambient air pollution were impacted by both the amount and the type of measurement error present, and these impacts varied substantially across pollutants.

Dominici et al. (2010) underlined that the issue of exposure error, if it is relevant in single pollutant approach, it takes on a larger role in multipollutant approach, where each type of measurement error can affect the exposure estimate for various pollutants differently. Among correlated pollutants, measurement error can affect the parameters in a multivariate regression model in different ways (Greenland 1980), and the direction of the bias depends on the sign of correlation between pollutants (Zeka and Schwartz 2004).

2.2.2 Correlation among exposure metrics

Correlation is a main concern when working with high-dimensional data sets, leading to collinearity or, in the regression setting, to multicollinearity.

Multicollinearity or nonorthogonality is a near-linear dependency between two or more predictors, that leads to a degeneracy in the system of equations in a multiple linear model (Morlini 2006).

In context of nonparametric or semiparametric models, such as GAM, Buja et al. (1989) used the term concurvity to describe nonlinear dependencies. Within this context, concurvity occurs when a function of one of the predictors can be approximated by a linear combination of function of the remaining covariates, with these functions being estimated in the same way as the corresponding functions in the original model (Ramsay et al. 2003). For example, in context of standard regression models applied to multipollutant data, the function might be the smooth function used to model the effects of confounding variables.

High degree of multicollinearity or concurvity has destructive effects on traditional multiple regression models and the validity of parameter estimation becomes questionable (Pitard and Viel 1997; Ramsay et al. 2003; Shieh and Fouladi 2003; Bagheri et al. 2010; MacLehose et al. 2007). Although it does not reduce the predictive power or reliability of the model, it causes unstable coefficient

estimation.

In traditional regression models, parameter estimation is a key part of model fitting and interpretation. Models are in this case used for hypothesis testing, probing the statistical significance of the effect of predictors on the response. The regression coefficients are actually partial coefficients, indicating the effect of unit changes in each predictor variable on the response variable, while holding all other predictor variables constant. The simultaneous regression of a response variable on correlated predictors changes the magnitude of the corresponding partial coefficient (Chatterjee et al. 2000). This because a high degree of multicollinearity or concurvity leads to high standard errors in the estimated coefficients and these inflated errors result in reduced statistical power to detect reliable effects of correlated variables, i.e., significance tests with inflated type 1 error (Ramsay et al. 2003).

Multicollinearity exists for several reasons. Most commonly, multicollinearity is intrinsic, meaning that collinear variables are different manifestations of the same underlying, and in some cases, immeasurable process (or latent variable). This can be the case with atmospheric multipollutant data that present, in fact, an intrinsic correlated nature. Collinearity also arises because of study design or model formulation. Montgomery et al. (2001) discussed that multicollinearity may be due to: (i) the data collection method employed, (ii) constraints on the model or as adding polynomial terms to the regression model, and (iii) an over-parametrization having more predictor variables than observations. Moreover, Kamruzzaman and Imon (2002) pointed out that high leverage points, namely observations that not only deviated from the same regression line as the other data but also that fall far from the majority of explanatory variables in the data set (Hocking and Pendelton 1983; Moller et al. 2005), can be source of severe multicollinearity.

In terms of multicollinearity detection, because this is a problem which exists in a data set, there is no statistical test for its presence (Bagheri et al. 2010). However, some diagnostic methods can be used to indicate the existence and extent of

multicollinearity in a data set. Most popular diagnostic tools of multicollinearity include the following examinations:

- the correlation matrix of predictors;
- the eigenvalues of $\mathbf{x}'\mathbf{x}$ (when the predictors are orthogonal or uncorrelated, all eigenvalues of the design matrix are equal to one and the design matrix is full rank; if at least one eigenvalue is different from one, especially when equal to zero or near zero, then nonorthogonality exists);
- the tolerance value or variance inflation factor (VIF) that measures how much the variance of the estimated regression coefficients are inflated as compared with the situation when the predictor variables are not linearly related (Marquardt (1970); generally, when VIF ≥ 10 then there is a problem with multicollinearity);
- the condition number (CN; Belsley et al. 1980) of the $\mathbf{x}'\mathbf{x}$ matrix that can be computed as the square root of the largest eigenvalue divided by the smallest eigenvalue (when CN is equal to one, the predictors are said to be orthogonal).

2.3 Methods for characterising air pollutant exposure metrics

Different methods have been used to characterise air pollution metrics with the aim to improve exposure assessment estimates to be included in epidemiologic analyses. In the next sections, a picture of the most applied statistically-based approaches within a time series design is presented, with emphasis on statistical methods for dealing with multipollutant metrics.

Overviews over methods and recent developments in analysing complex exposure metrics used in air pollution health studies can be found in Pitard and Viel (1997); Dominici (2004); Dominici et al. (2010); Billionnet et al. (2012); Sun et al. (2013) and Oakes et al. (2014).

2.3.1 Variable selection

A typical approach in dealing with different pollution metrics involves performing a multivariate regression model including these pollutants as predictors, and estimating the health effects of every pollutant while adjusting for the concentration of the additional pollutants. However, given the potential for a multicollinearity problem, it is common practice to choose a subset of these predictors to produce an optimal model.

The covariate selection could be performed without any statistical procedure, by applying knowledge of the atmospheric processes and/or excluding copollutants that are already known be highly correlated with a pollutant of interest (e.g., Ghosh et al. 2010). In most of the cases, however, variable selection methods are implemented statistically and the subset of covariates is chosen according some threshold for significance. In a Bayesian framework the problem is not longer searching for a single optimal model, but rather to attempt to estimate the posterior probability of all models within the considered class of models (O'Hara and Sillanpää 2009).

Automatic variable selection

Automatic variable selection, based on hypothesis tests and greedy algorithms like forward and backward stepwise selection, represents a traditional frequentist approach in dealing with multiple covariates. These methods are based on accepting the null hypothesis when covariates are non significantly associated with the outcome, and ignore the association between exposure of interest and covariates when deciding whether a given covariate confounds the association between exposure and outcome. When the candidate predictors are highly correlated, leading to dependence among tests, these techniques might produce results that are difficult to interpret and, eventually, produce biased point and interval estimates (Pitard and Viel 1997; Thomas et al. 2007b). For example, an estimated association could occur because a pollutant being analysed is a proxy for another

or for a mixture of air pollutants. Furthermore, these procedures do not always lead to the same model, and the set of selected covariates could change with samples drawn from the same population (Pitard and Viel 1997).

Deletion/Substitution/Addition technique

The Deletion/Substitution/Addition (DSA) algorithm has been presented as alternative model selection procedure for nuisance parameters. The technique was initially proposed for high-dimensional genomic data by Sinisi and van der Laan (2004), and successively was adopted in environmental science by Mortimer et al. (2008). This algorithm builds a space of candidate models based on so-called deletion, substitution and addition moves and utilises a loss function-based estimation procedure to distinguish between different models with respect to model fit. Briefly, the DSA procedure generates predictors as linear combinations of tensor product polynomial basis functions under user-specified constraints and progressively builds more complex models that contain more variables and interactions between them in an attempt to find a model that fits the data well and has good predictive performance. The algorithm performs a data-adaptive estimation through estimator selection based on cross-validation and the L_2 ("squared error") loss function. Thus, thanks to this feature of selecting models based partly on cross-validation, it avoids the problem of over-fitting data (i.e., it protects against selecting a too complex model). Compared with stepwise model selection procedures, the DSA algorithm presents some methodological advantages, as it is less sensitive to outliers (via the use of cross-validation during the search), and it allows the search to move among statistical models that are not nested (Dominici et al. 2008; Billionnet et al. 2012).

However, DSA approach, like any automatic variable selection technique, chooses a model based on a given data set and then estimates health effects in the same data assuming that the chosen model is correct. This could lead to misleading inferences (Dominici et al. 2003), with inflated effects and excessively optimistic estimates of precision (Benjamini and Yekutieli 2005; Thomas et al. 2007b;

Dominici et al. 2008). This because, if only a single "best" regression model is reported, the variance estimates for its coefficients do not fully reflect their uncertainties.

2.3.2 Bayesian model averaging

Typically time series studies report the health effect estimates from a single model, after extensive model selection and sensitivity analysis. Bayesian model averaging (BMA) has been proposed as a method to account for uncertainty in models with different covariates by combining inferences from a set of candidate models (e.g., Draper 1995; Hoeting et al. 1999). This procedure assigns probabilities or weights to each candidate model that reflect the degree to which the model is supported by the data. These probabilities can be used to produce weighted average estimates of the association between pollutants and health outcome, incorporating thereby information from each candidate model (e.g., Clyde 2000; Koop and Tole 2004; Martin and Roberts 2006; Chuang et al. 2010). Briefly, BMA can be formulated as follows.

Let $\mathcal{D}_n$ denote the data (given by the outcome response and the matrix of the predictors) and $M_1, \ldots, M_K$ the models considered. BMA estimates models for all possible combinations of the predictors in analysis and constructs a weighted average over all of them. If the matrix of the predictors contains P potential variables, thus 2^P variable combinations are estimated, that means 2^P models.

The likelihood function for M_k is $p(\mathcal{D}_n|\Theta_k, M_k)$, while the prior probability that M_k is the true model is $p(M_k)$. The integrated likelihood, that is the probability density of the data, conditional on the model M_k, is given by:

$$p(\mathcal{D}_n|M_k) = \int p(\mathcal{D}_n|\Theta_k, M_k)p(\Theta_k|M_k)d\Theta_k. \tag{2.12}$$

Thus, by Bayes's theorem, the posterior model probability of M_k is:

$$p(M_k|\mathcal{D}_n) = \frac{p(\mathcal{D}_n|M_k)p(M_k)}{\sum_{\ell=1}^{K} p(\mathcal{D}_n|M_\ell)p(M_\ell)} \tag{2.13}$$

BMA obtains the posterior inclusion probability of a candidate predictor by summing the posterior model probabilities across the models that include the predictor.

Now, let θ denote a specific quantity of interest (e.g., the relative risk associated with a particular increment in the air pollutant concentrations on a health outcome). The posterior distribution of θ, given the data, $\mathcal{D}_n$, is (Hoeting et al. 1999):

$$p(\theta|\mathcal{D}_n) = \sum_{k=1}^{K} p(\theta|\mathcal{D}_n, M_k)p(M_k|\mathcal{D}_n) \tag{2.14}$$

The first term on the right hand side of this equation is the posterior distribution of θ given a specific model M_k and the second term in the equation is the already seen posterior probability of the model M_k. In summary, this is the average of posterior predictive distribution for θ under each model considered, weighted by the corresponding posterior model probability.

This traditional BMA, despite its attractive qualities, can face several drawbacks in effect estimation (Thomas et al. 2007a; Wang et al. 2012). In fact, the regression coefficients may have different interpretation across models (i.e., different interpretations for individual pollutant); moreover this model can present a problem of overfitting in the context of confounders.

Wang et al. (2012) developed a Bayesian solution to adjustment uncertainty, called Bayesian adjustment for confounding (BAC). This approach is based on specifying two models: (i) the outcome as a function of the exposure and the potential confounders (the outcome model); and (ii) the exposure as a function of the potential confounders (the exposure model). The key of the approach of Wang and colleagues is the specification of a prior distribution such that, conditional on a predictor's inclusion in the exposure model, the same predictor should also have a higher probability to be included in the outcome model. The prior specification includes a dependence parameter, w, denoting the prior odds of including a predictor in the outcome model, given that the same predictor is in the exposure model. In the absence of dependence ($w = 1$), BAC reduces to

traditional BMA.

Martin and Roberts (2006) implemented model averaging using a bootstrap-based procedure, showing that it is competitive with BMA in time series studies of PM and mortality. Subsequently, Roberts and Martin (2010) proposed an extension of this method through a double bootstrap, showing an increased performance attributable to a reduction in the variance of the estimates.

2.3.3 Hierarchical models

Hierarchical models, also known as multi-level models (e.g., Gelman and Hill 2007) have been extensively used in environmental statistics and epidemiology to construct spatial, temporal and spatio-temporal exposure modelling and to study health effects of air pollution.

Berliner (1996) provided a conceptual definition of a hierarchical model in a Bayesian framework, widely adopted by the scientific community, see for example Cressie and Wikle (2011) and Royle and Dorazio (2008). The skeleton for hierarchical modelling is constituted by three entities: the data, the process and the parameters. The data model expresses the conditional distribution of data given both the process (this hidden process is the focus of inference) and the parameters. The process models the uncertainty in the hidden true process through a probability distribution on the phenomenon in study (essentially it describes the dynamic of the process). The parameter model presents additional structures to relate the parameters of the observations and the process parameters. Within this framework, Bayes theorem can then be used to obtain the posterior distribution of the process and parameters updated by the data, in the following way (Wickle 2003):

$$p(\text{process, parameters} \mid \text{data}) \propto$$
$$p(\text{data} \mid \text{process, parameters})\ p(\text{process} \mid \text{parameters})\ p(\text{parameters}).$$

The modelling approach presented in chapter 3 is based on this framework.

The Health Effect Institute (HEI 2010) has emphasized the benefits of a Bayesian hierarchical model in environmental studies, that can be summarised in the following headings: (i) modular model elaboration, (ii) integration of different sources of information, (iii) coherent propagation of uncertainty, (iv) borrowing of strength, and (v) integrated treatment of information at different levels.

In air pollution time series studies, Bayesian hierarchical models have been the subject of increasing attention in both single as well as multiple pollutant approaches (e.g., Shaddick and Wakefield 2002; Huerta et al. 2004; Sahu et al. 2006; Cocchi et al. 2007; Lee and Shaddick 2010; Chang et al. 2011; Cameletti et al. 2011).

Recently, this approach has been used in multi-site (or multi-city) time series studies, performed with the aim to introduce a spatial dimension in the estimation of the short-term health effects of outdoor pollution over a region in study. Here, city-specific data on air pollution and health are collected under a common framework, and subsequently analysed using a uniform statistical approach (e.g., Dominici et al. 2000; Huang et al. 2005; Peng et al. 2005). So far, in multi-site time series studies, hierarchical models represent the statistical framework for summarising health risks associated to air pollution through different cites. Typically, these multi-site studies present a multi-stage structure, where at the first stage the association between pollutant(s) and health is assessed using Poisson time series regression models (i.e., GLMs or GAMs) at city level, controlling for trend, season and meteorological variables, and at the second stage the results from multiple sites are combined by assuming that the true city-specific estimates have a common mean (called the pooled relative risk) and variance that reflect the variability across cities of the true estimates (called the heterogeneity parameter) (Daniels et al. 2004). The application of Bayesian hierarchical multi-pollutant models, within a multi-site time series framework, given by Peng et al. (2009) and Bell et al. (2009), provided evidence that the chemical composition of fine particle air pollution affects its toxicity. In terms of adverse health effects, (i) Peng et al. (2009) pointed out that the ambient concentrations of EC

and OC were associated with the largest risks of emergency hospitalisation across the major chemical constituents of $PM_{2.5}$; (ii) Bell et al. (2009) showed that in communities and seasonal periods in which $PM_{2.5}$ had higher fractions of nickel, vanadium, and EC and/or their related sources, the risk of hospital admissions was higher.

Recently, Bobb et al. (2013) observed, however, that when the goal is to estimate the health effects of many pollutants jointly, a straightforward application of Bayesian hierarchical models can be challenged by the need to specify a random-effect distribution on a high-dimensional vector of nuisance parameters, which often do not have an easy interpretation. To overcome this issue, Bobb et al. (2013) introduced a reduced Bayesian hierarchical model, based on an integrated likelihood for summarising information about the main parameters of interest.

Bayesian hierarchical models are largely used in environmental epidemiologic studies to combine data from different sources. In this context, they have grown in popularity in modelling situation characterised by misaligned data, that is the case of observations collected at different spatial locations or spatial resolutions. In time series studies this represents one of the potential sources of measurement error, as discussed in section 2.2.1. As already seen, this may be due to the misalignment between pollutant concentrations measured at monitor stations and health outcomes averaged over the study area. Additionally, it may occur when exposure metrics are collected at different spatial resolutions, such as ambient data measured from an available network of fixed monitoring stations and output from atmospheric deterministic models supplied for grid cell. These misalignments are known as *change of support problem* (Gelfand et al. 2001; Gotway and Young 2002).

The literature offers a number of studies developed within the hierarchical modelling framework facing this problem. Examples are provided by Choi et al. (2009); Peng and Bell (2010) and Lee and Shaddick (2010). A discussion of hierarchical statistical models for exposure data collected over different spatial

scales is provided in the background section of chapter 3.

2.3.4 Air pollution indices

Composite air quality indicators have been proposed in literature to represent a complex pollution scenario. Hong et al. (1999) developed a combined index of PM_{10}, NO_2, SO_2, and CO, by the sum of a 5-day moving average divided by the mean values of each pollutant. The score was selected a priori and an equal weight was assigned to each pollutant included in the index, thus loosing the possibility to understand the effect of each type single pollutant (Billionnet et al. 2012; Roberts 2006).

To overcame this limit, Roberts (2006) and Roberts and Martin (2006a) introduced a weighted model for disentangling the joint effects of multiple air pollutants. Thus, time series data were used to assign each air pollutant a weight indicating the pollutant's contribution to the air pollution mixture, with a constraint on the weights to be non-negative and scaled to sum to one.

Bruno and Cocchi (2002) described a methodology to build indices from data collected from multiple monitoring sites, via a hierarchical aggregation based on successive selection of order statistics (i.e., on percentiles and on maxima).

Lee et al. (2011) proposed a Bayesian geostatistical modelling approach that allowed the construction of intervals of uncertainty for a composite index based on four pollutants, CO, NO_2, O_3 and PM_{10}. Powell and Lee (2014) extended this approach within a three-stage Bayesian hierarchical framework, which comprised: (i) a geostatistical model to estimate the posterior predictive distribution of a spatially representative measure of a single pollutant, (ii) the combination of these distributions across pollutants to produce an air quality indicator, and (iii) a model for estimating the health effect of either a single pollutant and the composite air quality indicator.

2.3.5 Shrinkage methods

Shrinkage methods techniques include penalized regression such as ridge regression (Hoerl and Kennard 1970), Least Absolute Shrinkage and Selection Operator (lasso) technique (Tibshirani 1996) and partial least squares regression (Sjöström et al. 1983). To address the multicollinearity, these methods use a penalty term to deliberately bias, or shrink, their coefficient estimates to account for excessive variation in the original (unbiased) estimates.

In regularized linear regression, two main constraints have been used, consisting in: (i) fixing the upper bound of the L_1 norm (i.e., the sum of the absolute values) of the vector of regression coefficients, and (ii) specifying the upper bound for the L_2 norm (i.e., the sum of the squares). Ridge regression uses a L_2 penalization in high dimensional problems. It does not perform variable selection, it only shrinkages toward zero. The model replaces the standard least square estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \tag{2.15}$$

with

$$\hat{\boldsymbol{\beta}}^{c} = (\mathbf{x}'\mathbf{x} + c\mathbf{I})^{-1}\mathbf{x}'\mathbf{y} \tag{2.16}$$

where c is the ridge parameter and I is the identity matrix. If $c = 0$, thus $\hat{\boldsymbol{\beta}}^{0}$ is the usual least square estimator. The c is also referred to as the biasing parameter. When c increases from 0, the estimator $\hat{\boldsymbol{\beta}}^{c}$ becomes biased, but the variance decreases (Pitard and Viel 1997). Ridge regression, however, does not produce a parsimonious model, as all predictors are kept in the model.

To achieve sparsity (i.e., favoring null regression coefficients), the lasso technique is more appropriate. It uses the L_1 penalty on the regression coefficients and performs shrinkage and variable selection simultaneously. It aims to minimize $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{x}'\mathbf{x}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ with $\boldsymbol{\beta}$ subject to $\sum_{j=1}^{P} |\beta_p| \leq s$ with s a user-specified parameter that controls the amount of shrinkage. However, it has been shown that in the presence of highly correlated covariates, lasso tends to select only one variable

among them (Zou and Hastie 2005).

Roberts and Martin (2005) provided a critical assessment of shrinkage-based regression methods in comparison to the standard Poisson log-linear model, to estimating the adverse health effects of multiple air pollutants. The authors concluded that, although ridge regression produces more accurate estimates than the lasso, the latter produces more interpretable models. In any case, both these techniques provide more accurate estimation of pollutant effects than that provided by the standard model.

A compromise between ridge regression and lasso, achieving both shrinkage and automatic variable selection, was given by Zou and Hastie (2005) that proposed the "elastic net" criteria. The elastic net procedure uses a penalty, λ, that is a convex combination of L_1 and L_2 norms of the regression coefficients where the two extreme cases of $\lambda = 0$ and $\lambda = 1$ correspond to the lasso and ridge regression constraints, respectively. Elastic net is considered to outperform lasso regression by encouraging grouping effects (i.e., achieving shrinkage on block of covariates, that is some blocks of regression coefficients are exactly zero) and improving predictions (Zou and Hastie 2005).

In a recent review Sun et al. (2013) positively stressed these methodologies in assessing the health effects of pollution mixtures as deserving further investigation.

Partial least-square, finally, creates orthogonal score vectors (also called latent vectors or components) as linear combinations of the original regression variables and represents a compromise between maximizing the explained variance of the predictors and maximizing the correlation between the predictors and the outcome.

2.3.6 Feature extraction

The goal here is to reduce a large number of predictors to meaningful summaries that might be used in further analyses and identify underlying structures. Traditional methods include factor analysis (FA) and principal components ana-

lysis (PCA) (Burnett et al. 2000; Cox 2000). The main idea is to transform the original feature space P into a space in which the data are not correlated (that is, the variance of the data is a maximum). FA seeks linear combinations of unobserved variables, called factors, that represent underlying fundamental quantities of which the observed variables are expressions. FA and PCA can lead to the same results, even though they are not identical methods. In general, FA is a model for the correlation structure, plus measurement errors; while PCA uses the covariance structure of the data which is expanded in an ordered set of components of decreasing variance. Both these approaches are extensively used also as source apportionment methods (see section 2.3.7). Here, it is deeper described the PCA methodology, while FA is better stressed in the next section.

The goal of PCA is to describe the variation of a set of multivariate data in terms of a set of uncorrelated new, latent variables, called principal components, which are obtained as linear combinations of the original variables. Each principal component is a weighted average of the underlying indicators. Weights are chosen so as to maximize the explained proportion of the variance in the original data. This is obtained by computing the eigenvalues and the eigenvectors of the covariance matrix of the initial data and selecting the eigenvectors that have the largest eigenvalues. These component will represents the axes of the new transformed space. Formally, consider P pollutants, $x_1, x_2, \ldots, x_P$. The first principal component, z_1, given by the linear combination of the original variables, is:

$$z_1 = a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,P}x_P \tag{2.17}$$

with constraint $a_{1,1}^2 + a_{1,2}^2 + \cdots + a_{1,P}^2 = 1$. The second principal component z_2 is computed in the same way, but it is uncorrelated with the first principal component, and it accounts for the next highest variance:

$$z_2 = a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,P}x_P. \tag{2.18}$$

The remaining principal components are chosen in the same way. In synthesis, the

transformation of the original variables to principal components, using a matrix notation, is: Z = AX, accounting for the variance in decreasing proportions. The rows of the matrix A are called the eigenvectors of the matrix Σ_x, that is the variance-covariance matrix of the original data. The elements of the eigenvector are the weights a_{ij} (also called as loadings, so the matrix A is also called a loading matrix). The elements in the diagonal of matrix Σ_x are known as the eigenvalues. These eigenvalues are the variance of the principal components (i.e., the first eigenvalue is the variance of the first component, the second eigenvalue is the variance of the second principal component, and so on).

A central question arises is how many components are needed to provide an adequate summary of the original data. Several criteria have been proposed for determining how many principal components should be investigated (Everitt and Dunn 2001). The most common are: (i) include the components that explain a relatively large percentage of the total variation (e.g., between 70 and 90 per cent), (ii) choose the principal components with eigenvalues over 1 when the correlation matrix is used or less than the average variance explained when a covariance matrix is used, (iii) use the scree plot of the eigenvalues, that can indicate an obvious cut-off between large and small eigenvalues.

Solutions from PCA are largely used for characterising multiple air pollution exposure metrics and often they are combined with regression analysis (e.g., Zhao et al. 2011).

Despite this, PCA presents some drawbacks. The components which explain the major proportion of the variance of covariates are not necessarily those most correlated to the outcome and the results can be ambiguous and difficult to interpret (Pitard and Viel 1997), since the basic profile may include many negative values. Moreover, the physically meaningful representation can be found only after a series of transformations called rotations.

Supervised principal component analysis (SPCA) (Bair et al. 2006) represents a modified version of PCA for use in regression problems in which the number of predictors greatly exceeds the number of observations. Roberts and Martin

(2006b) present an implementation of SPCA to characterise exposure to multiple pollutants. For this purpose, SPCA is similar to PCA except that it uses a subset of the multiple pollutants that are selected on the basis of their association with the adverse health outcome of interest, rather than only on intrinsic properties of the covariate space.

In a Bayesian framework of latent factor models of time series (e.g., Aguilar et al. 1998), Reich et al. (2009) proposed a supervised dynamic factor model with the aim of analysing a time series of particles with diameter less than 0.40 μm in 17 different size bins, and to relate the various PM diameters with non-accidental mortality. Because the particles present a natural ordering of diameters, the authors proposed an extension of the usual latent factor model, taking into account the similarity between adjacent particle diameters.

2.3.7 Source apportionment

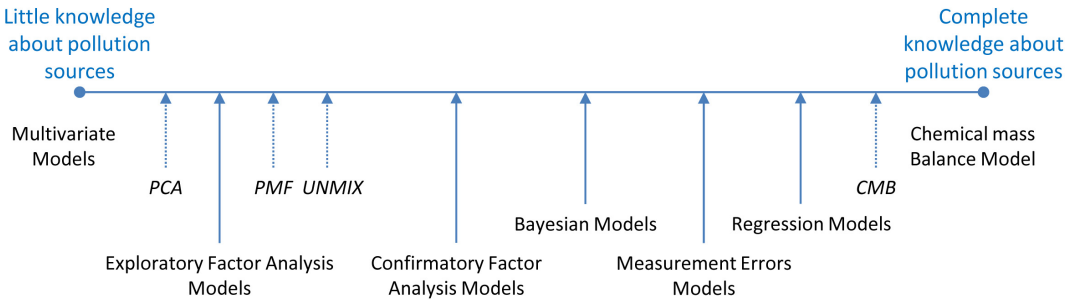
Source apportionment defines a class of methods aimed at partitioning pollution to the sources from which they were emitted. Has been hypothesised that similar pollutants, originating from different sources, may have different levels of toxicity (Christensen et al. 2006). Therefore, epidemiologic studies that incorporate a quantification of the impact of various pollution sources may provide a step toward targeting important causal agents, and support the development of effective control strategies. Moreover the quantification of the health risks associated with sources, rather than single pollutants, may also capture complex interactions that more accurately reflect the etiologic relationships between particles and adverse health outcomes (Sarnat et al. 2008; Stanek et al. 2011b).

Currently, two basic approaches have been generally used for source apportionment: dispersion modelling and receptor modelling. The former relies on mathematical description of physical-chemical process taking place in atmosphere based on emission inventories and simulates the aerosol formation, transport and deposition, so as to calculate the concentrations at various locations; the latter is based on statistical procedures for identifying and quantifying the sources of

pollutants on the basis of mixture of chemicals measured at a given receptor site. Compared with dispersion models, receptor models do not require detailed emission inventories and source locations (that are not always available), and in the last three decades they have been extensively used (Viana et al. 2008), especially for investigating the roles of PM components and sources (e.g., Hopke et al. 2006; Ito et al. 2006; Mar et al. 2006; Sarnat et al. 2008; Belis et al. 2013).

Schauer et al. (2006) illustrated the variety of approaches for estimating pollution source contributions using receptor models along a continuum of a priori knowledge about the sources, with multivariate models and chemical mass balance (CMB) as the two main extremes. Figure 2.3 shows a modified version of the graphical representation by Schauer and colleagues.

Figure 2.3: Approaches for estimating pollution source contribution using receptor models; specific models are showed in italics and with dotted arrows (modified from Schauer et al. (2006)).



In particular, if the sources are known and detailed information on source profiles at a given site is available, the most used technique is represented by the CMB model (Watson and Schoen 1984; Chow and Watson 2002); but when little or nothing is known about the nature of the pollution sources, exploratory FA (e.g., Thurston and Spengler 1985), confirmatory FA (Christensen and Sain 2002; Christensen et al. 2006), PCA (Thurston and Spengler 1985), absolute principal components analysis (APCA; Thurston and Spengler 1985), UNMIX (Henry and Norris 2002), positive matrix factorization (PMF; Paatero and Tapper 1994; Kim et al. 2003), Bayesian analysis (Park et al. 2001, 2002; Lingwall et al. 2008; Heaton et al. 2010) and measurement error modelling (Watson and Schoen 1984)

are preferred.

In general, the principle on which receptor models are based is the chemical mass conservation law, that essentially states the mass conservation between emission sources and receptor site (Pollice 2011). This implies that the concentrations of a specific pollutant, such as PM, at receptor site are a linear combination of all responsible sources. Therefore, given $j = 1, 2, \ldots, P$ chemical species in the $t = 1, 2, \ldots, T$ samples as contribution from K independent sources, the chemical mass balance equation can be written as:

$$y_{tj} = \sum_{k=1}^{K} g_{tk} f_{kj} \tag{2.19}$$

where y_{tj} is the jth elemental concentration measured in the tth sample, g_{tk} is the contribution of the kth source to the tth sample, f_{kj} is the concentration of the jth species in material emitted from each source k.

CMB model uses the chemical and physical characteristics of gases and particles measured at both the source and the receptor site to identify the presence of pollutants and to quantify the contributions of the source. In this case, because source profile, f_{kj}, are known, the source contributions, g_{tk}, can be determined from the linear regression of the y_{tj} on f_{kj} (Watson and Schoen 1984). This model performs well when changes of the source profiles between the source and the receptor may be considered minimal. However these requirements are almost never completely fulfilled. Moreover, CMB is constrained by the need to include secondary aerosols not as components of emission source profiles but as specific chemical compounds. This is often regarded as a limitation (Viana et al. 2008).

Traditional FA has been largely used to identify and quantify sources and their impact over a set of samples. The chemical mass balance equation holds also for FA, however in this case g_{tk} and f_{kj} are derived by the FA from the correlation matrix and are output of the FA. These models present the advantage that they can identify and quantify nontraditional aerosol like secondary aerosol and can incorporate non-PM tracers such as the gaseous pollutants (Thurston et al. 2005).

Also PCA (e.g., Roscoe et al. 1982) and APCA have been extensively used in the past to produce quantitative source apportionment, however more recently multivariate factor analysis tool like PMF and UNMIX are gaining relevance for their increased flexibility. Both these techniques place restrictions on the possible source impact solution to require them to meet certain physical constraint (Hopke et al. 2006).

In particular PMF, has recently encountered a large use. This technique was developed in order to resolve some limitations of the standard techniques such as FA and PCA. In the line with these, the goal of PMF is to explain the observed data using a limited number of basic components, which approximate the original data as accurately as possible. However, PMF constrains the source profiles and source contributions to be non-negative to match the physical reality of the problem. By requiring non-negativity, PMF is able to produce results (i.e., the matrix factors) which are easier to interpret. Another aspect of PMF is the optimal use of the error estimates. It, in fact, computes the solution by minimizing the least squares error of the fit, weighted with the error estimates. The general equation seen before, for PMF becomes:

$$y_{tj} = \sum_{k=1}^{K} g_{tk} f_{kj} + e_{tj} \tag{2.20}$$

where e_{tj} is the residual for each sample/species. The objective is to find g_{tk} and f_{kj} by minimizing the residual error e_{tj}. For this a weighted least square approach is used. It involves minimization of an objective function, Q, given as:

$$\text{minimize } Q = \sum_{t=1}^{T} \sum_{j=1}^{P} \frac{e_{tj}^2}{s_{tj}^2}$$

$$\text{subject to } g_{tk} \geq 0,\ f_{kj} \geq 0$$

where s_{tj} is an uncertainty estimate in the jth species measured in the tth sample.

Dominici et al. (2010) underlined the attractiveness of the source apportionment methods from a regulatory standpoint because they help to identify specific

targets of regulatory intervention. However, the same authors, discussed also the following limitations: (i) the focus on few sources might lead to the omission of other important sources; (ii) some source-related approaches could require information that may be poor quality or not available at all; (iii) the results are not easily generalisable because of the location-specific nature of most of the sources; and finally (iv) these methods substitute one complex mixture (the air) with another complex mixture (the source).

2.3.8 Clustering

Cluster analysis, or unsupervised classification in machine learning, aims to identify homogeneous groups or clusters in the data of previously unknown structure, so that objects within a cluster have similarity in comparison to one another, but are dissimilar to objects in other clusters.

There is a very large number of clustering methods. Broadly speaking, these methods fall into two types: partitional (or flat) and hierarchical approaches (Murphy 2012). Partitional clustering algorithms find all the clusters simultaneously as a partition of the data, and these partitions are independent of each other. Hierarchical clustering algorithms recursively create nested tree.

A literature review of clustering specifically applied on time series data (in different fields) may be found in Liao (2005) and Kavitha and Punithavalli (2010). This section provides a description of commonly used clustering techniques in air pollution science, starting from traditional techniques to more sophisticated model-based methods in which clustering is defined in a probabilistic framework.

Model-free methods

Conventional clustering algorithms applied in air pollution time series studies are based on heuristic criteria and not on formal models.

Hierarchical clustering

Hierarchical clustering (Johnson 1967) is a recursive partitioning process, which results in a hierarchical nested cluster structure. The partitions can be visualized using a tree structure, where each level of the resulting tree is a segmentation of the data. A hierarchical clustering algorithm can be agglomerative or divisive. The first one works starting with each data point in its own cluster, then merges the most similar pair of clusters successively to form a cluster hierarchy; the second one works starting with all the data points in one cluster, and recursively divides the cluster into smaller clusters. At each stage of hierarchical clustering, the splitting or merging is chosen so as to optimize some criterion.

Conventional agglomerative hierarchical methods use single linkage (minimum distance between groups), complete linkage (maximum distance between groups), or average group (average distance between groups). Hierarchical algorithms present some attractive features, as they do not require prior specification of the number of clusters and allow the data to be view at many levels of granularity, all at the same time. However, hierarchical classification is very time-consuming and, mostly, at each step, the partitioning criterion is not global but depends on the classes already obtained (Billionnet et al. 2012).

K-means

K-means (Hartigan and Wong 1979) is a popular hard partitional clustering algorithm, in which each point is assigned to only one particular cluster. It is applied in situations in which the variables in analysis are quantitative and uses typically the squared Euclidean distance (or a weighted Euclidean distance) to measure dissimilarity. The objective of the K-means algorithm is to minimize a cost function given by the sum of squared distances between all points and their cluster centres (centroids or means). In particular, given a set of T data points $x_1, \ldots, x_T \in \mathbb{R}^d$ and a fixed number K of clusters, K-means attempts to minimize

the following clustering objective function:

$$Q(c_1, \ldots, c_K) = \frac{1}{T} \sum_{t=1}^{T} \min_{k=1,\ldots,K} \parallel x_t - c_k \parallel^2 \tag{2.21}$$

where $c_1, \ldots, c_K$ denote the centers of the K clusters. The algorithm needs to be randomly initialised specifying the initial cluster centers and works by iteratively assigning points to the nearest centroid, then repositioning those centroids to the mean of the points. The process terminates when a convergence condition is satisfied. Specifically, the main steps of K-means algorithm are as follows (Jain 2010):

1. Select an initial partition with K clusters.

2. Generate a new partition by assigning each pattern to its closest cluster center.

3. Use the classification from step 2 and compute the vector of new cluster centers.

4. Repeat steps 2 and 3 until cluster membership stabilizes.

During this procedure the number of clusters can not change.

Despite its popularity K-means clustering algorithm presents several well known drawbacks. It is extremely sensitive to cluster center initialization and outliers, as it is based on the mean, that is a descriptive statistic not robust to outliers (Billionnet et al. 2012). Moreover, the results depend on the order of the objects in the input file (Kaufman and Rousseeuw 1990).

An alternative to K-means is represented by *Partition Around Medoids* (PAM) technique (Kaufman and Rousseeuw 1990). PAM partitions the data identifying clusters by the medoids, which are robust representations of the cluster centers that are less sensitive to outliers than other cluster profiles, such as the cluster means of K-means. The algorithm works by selecting the first medoid, choosing the observation for which the sum of dissimilarities between it and all other observations (i.e., the sum of the distances between each observation and the

proposed medoid) is minimized. The second medoid is selected in much the same manner, and so forth.

Temporal clustering analyses have been successfully applied in air pollution exposure assessment, involving both agglomerative hierarchical clustering (Gu et al. 2012) as well as K-means partitioning clustering (Austin et al. 2012). Recently, K-means clustering solutions of air pollutants have also been used as covariates within health model effect estimation (Matyasovszky et al. 2011; Zanobetti et al. 2014). An interesting strategy that combines functional data analysis (that is, data analysis with curves) with clustering is provided by Ignaccolo et al. (2008), which converted discrete time series into functional data through the estimation of spline coefficients, and then partitioning the estimated coefficients by PAM classification.

These traditional clustering methods are based on similarity/dissimilarity measures between objects that are essentially described in terms of distance (e.g., Euclidean distance), they require that the time series of each pollutant has exactly the same dimensionality (i.e., they do not allow the inclusion of records which have missing data). Further, as underlined by Liao (2005), the distance-based clustering methods can not be easily extended to time series data, where an appropriate distance-measure is rather difficult to define. Finally, and most importantly, these techniques do not allow an assessment of the statistical properties of the solutions provided, for example they do not provide an assessment of clustering uncertainties.

Model-based methods

Mixture models provide a model-based approach to clustering. These methods represent an alternative approach to heuristic techniques for clustering data, providing flexible probabilistic models of the data that are viewed as coming from a mixture of probability distributions, each representing a different group or cluster. They are based, in fact, on the idea that the data are clustered

using some assumed mixture modelling structure, specifically, that the data are generated by a mixture of underlying probability distributions in which each component represents a cluster. Mixture models comprise a finite or infinite number of components.

Finite mixture models

Probabilistic finite mixture model are well described in McLachlan and Peel (2000); Fraley and Raftery (2002); McLachlan and Baek (2010).

Let a data set of random values $\mathbf{x} = (x_1, \ldots, x_T)$ be drawn independently from some unknown distributions. In a finite mixture model, the probability density for $\mathbf{x}$ is modelled as a mixture of K component densities $p_k(\mathbf{x}|\boldsymbol{\theta}_k)$ on some unknown proportions $\pi_1, \ldots, \pi_K$, that is:

$$p(\mathbf{x}|\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}|\boldsymbol{\theta}_k) \tag{2.22}$$

where $\boldsymbol{\theta}_k$ is the set of parameters defining the kth component. In the clustering context, each component density in the above equation represents the distribution of a single cluster. $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the mixture probabilities (or mixing weights) which must be positive and sum to one. π_k represents the probability that an observation belongs to the kth component. $p_k(\cdot, \boldsymbol{\theta}_k)$ is a distribution belonging to a parametric family. Therefore, different choices for the component densities allow to model different type of data. For example, a Gaussian distribution might be used for continuous data, whereas a Multinomial might be used for categorical data.

To identify the cluster from which each observation is drawn, it is common in mixture modelling (e.g., Dempster et al. 1997) to introduce a latent component $\mathbf{z} = (z_1, \ldots, z_T)$, which can take values on $z_t \in (1, \ldots, K)$, such that $z_t = k$ if the tth observation is assigned to cluster k. The z_t's are assumed to be independently and identically distributed with probability mass function $p(z_t = k) = \pi_k$.

Various methods have been developed for estimating the parameters in finite

mixture models. Two widely used in practice are maximum likelihood (ML) method and Bayesian method. Estimation for model-based clustering carried out using ML estimation can be based on the expectation-maximization (EM) (Dempster et al. 1997) algorithm (e.g., Fraley and Raftery 1998; McLachlan and Peel 2000). Alternatively, a fully Bayesian approach can be used, placing a prior distribution on the parameters and computing a posterior distribution. The indicators or labels $\mathbf{z}$ can be assumed to be governed by a Discrete distribution (Multinomial) parameterised with the $\boldsymbol{\pi}$ vector of probabilities, then the most sensible choice for its prior is the Dirichlet distribution, that is commonly used to model a distribution over probabilities and is conjugate prior for the Multinomial:

$$\begin{aligned} z_t|\boldsymbol{\pi} &\sim \text{Mult}(\pi_1, \ldots, \pi_K) \\ \pi_1, \ldots, \pi_K &\sim \text{Dir}(\boldsymbol{\alpha}) \end{aligned} \tag{2.23}$$

where $\boldsymbol{\alpha}$ is the vector parameter of the Dirichlet distribution, that will be described in more details in chapter 4 (section 4.2). Computationally, these models rely on MCMC sampling (Neal 2000).

Frühwirth-Schnatter and Kaufmann (2008) showed that model-based clustering based on finite mixture models extends to time series in quite a natural way. In the air quality field, Gómez-Losada et al. (2014) applied a finite mixture model for characterising air pollution mixtures, using maximum likelihood, via the expectation-maximization algorithm.

Infinite mixture models

A long-standing issue that finite mixture models share with many traditional clustering methods (e.g., K-means), is the a priori determination of the number of clusters K. Different methods can be used to estimate the number of components (i.e., clusters), using for example model selection criteria. However, an alternative way to handle this problem is to adopt a Bayesian nonparametric modelling approach, where the number of mixture components is not fixed in advance, but is determined by the model and the data. These models can be implemented using

a Dirichlet process (DP) (Ferguson 1973; Antoniak 1974), a stochastic process commonly used in Bayesian nonparametrics to model the uncertainty about the functional form of the distribution of the parameters in a model. The support of the DP is restricted to discrete distributions and this results in a clustering effect that avoids the selection of a pre-defined number of clusters. This approach has been used in developing the study presented in chapter 4, and the statistical background is described in section 4.2.

Within the specific time series context, methodological applications of Bayesian nonparametric methods include Iorio et al. (2004); Müller et al. (2004); Griffin and Steel (2006); Dunson et al. (2007); Griffin and Steel (2011); Fox and Jordan (2013). However, at the moment not specific application in air pollution time series analysis with specific aim of data clustering have been proposed in literature.

2.4 Discussion

This chapter has reviewed the most common statistical methods used in time series studies of air pollution and health, with particular emphasis in the characterisation of the exposure metrics.

Most of the literature have focused on the health risk associated with single pollutants or sources one at a time. This is expressed, for example, by the studies on the total mass concentration of particles, performed without considering the heterogeneity in their chemical and physical composition. In the last decade, however, there has been a growing interest, in environmental research and air quality management communities, in assessing the health effects of simultaneous exposure to different air pollutants and sources. Therefore, this chapter has attempted to describe the main methodological challenges associated with the analysis of air pollution matrices to derive informative exposure to be used in time series studies. These mainly relate to the complex nature of PM components and sources, such as the correlation structure of the data, their variability over

space and time and the potential confounding effects of non-pollutant factors.

The main classes of statistical approaches to modelling these air pollutant metrics, within an ecological time series design, consist on regression-based methods, often associated with strategies of variable selection and dimension reduction techniques. Within the latter, factorial-based analysis and clustering are gaining a growing attentions.

The approaches presented have different strengths and weaknesses. None of these represents a gold standard for analysing multiple pollutant and source metrics, however they have different flexibility and, in particular, capability in accounting for the uncertainties involved.

The next two chapters of this thesis, provide a deeper understanding of the potentialities associated with two of the approaches here described, providing a concrete application in an urban polluted environment. In detail, a Bayesian hierarchical modelling approach is used to combine different sources of particles, taking into account the spatial and temporal dependency in the pollution data (chapter 3), and a Bayesian clustering approach is used to infer the health risk of different particle profiles (chapter 4).

3 | Hierarchical spatio-temporal modelling for airborne urban particulate

Work in this chapter aimed to provide a statistical approach for modelling estimation and prediction of PM in an urban environment for use in short-term health effect studies. It provides a strategy for combining multiple sources of PM data, integrating output from an atmospheric numerical model with measurements at monitoring stations, and evaluating as well the influence of a set of covariates with direct or indirect influence on the pollution distribution. The study was motivated by the need to improve urban exposure to estimate the short-term health effects of PM_{10} in London. Because the data set available to this purpose presented both a temporal and a spatial dimension, the preference went for a spatio-temporal approach. This offered the chance to explore exposure modelling that could overcome the issue of the spatial misalignment error in traditional time series design.

The chapter is organized as follows. Section 3.1 provides the literature background. Section 3.2 discusses the environmental and statistical perspectives adopted for the development of the models. Furthermore, section 3.3 reviews some statistical preliminaries for point-referenced process. Sections 3.4 and 3.5 describe the data along with the exploratory analyses executed. Section 3.6 presents all the aspects associated with the model development, including model formulation, computation, prediction and sensitivity analyses. Section 3.7 describes the results and section 3.8 provides a final discussion on the modelling approach here presented.

This chapter is based on a recently published peer-reviewed article in *Journal of Exposure Science and Environmental Epidemiology* by Pirani et al. (2014). The paper was coauthored by John Gulliver, Gary W. Fuller and Marta Blangiardo who provided data sets, supervised the analyses and contributed to the interpretation of the results.

3.1 Background

Air pollution time series analyses are typically performed using exposure data obtained from a single or few monitoring sites centrally located and assumed to be representative of the average exposure experienced by a community. Air pollutant metrics from these central-site monitors could, however, be lacking of spatial resolution and lead to a potential bias.

Several studies have pointed out that an accurate assessment of temporal and spatial variations in ambient air pollution concentrations is a critical point for the interpretation of time series epidemiologic studies (e.g., Sarnat et al. 2010; Peng and Bell 2010; Lee and Shaddick 2010; Bell et al. 2011; Shaddick et al. 2013). In order to improve health impact studies, enhanced spatial and temporal coverage and resolution is encouraged. This is especially relevant for city-wide exposure assessment because of the heterogeneity in emission sources and complex pollutant flows due to urban morphology (Denby et al. 2009). In this context, geographic information system (GIS)-based methods like land use regression models (LUR) (Briggs et al. 2000; Hoek et al. 2008) have been successfully applied to estimate long-term (e.g., annual) ambient concentrations (Gulliver et al. 2011; Tang et al. 2013), but these techniques are not appropriate for short-term modelling as they do not include the influence of both changing source emissions or meteorology.

To provide accurate estimates of air pollution concentrations to use in health effect studies, researchers are increasingly turning to *statistical* or *deterministic dispersion models*. The former approach typically considers series of data collected at monitoring sites and characterises these with spatial or spatio-temporal

structures; in this context the Bayesian paradigm has experienced a substantial increase in usage in the last decade (e.g., Shaddick and Wakefield 2002; Huerta et al. 2004; Sahu et al. 2006, 2007; Cocchi et al. 2007; Chang et al. 2011; Cameletti et al. 2011). The latter approach simulates the dispersion of air pollution concentrations through deterministic models based on complex mathematical description of physical-chemical processes taking place in the atmosphere. Because the measurements at ambient monitoring stations can be sparse and irregularly spaced, as well as affected by missing data, the use of deterministic dispersion models has become increasingly popular due to their more comprehensive coverage over space and time. However, deterministic models are affected by different sources of uncertainty when compared with measurements; the output depends not only on accurately characterising source emissions, meteorology and geographical features of the dispersion environment, but also on the model configuration options selected by the user (for instance, several numerical models present options to apply diurnal, weekly and monthly profiles to the emission sources). With respect to the issue of numerical uncertainty in deterministic models, a critical role is played by the ambient measurements as they are frequently used to develop, evaluate and calibrate the air quality models. The process of calibration is somewhat contentious but it is widely accepted that the use of measurements can lead to improved model performance where some inputs are not fully parameterised (National Research Council 2007).

Over recent years, approaches to tackle this problem have been embedded within a wider data assimilation framework. Wikle and Berliner (2007) define data assimilation as:

> ...an approach for fusing data (observations) with prior knowledge (e.g., mathematical representations of physical laws; model output) to obtain an estimate of the distribution of the true state of a process.

Data fusion for environmental exposure assessment describes an approach for

synthesize multiple data sources that are informative about spatial and spatio-temporal processes over to the same geographic domain and the same time period (Banerjee et al. 2014). Within this area of research, the Bayesian approach has provided a natural choice for combining information sources while managing their uncertainties.

Examples of data assimilation studies focused on spatial data include works by Wikle and Berliner (2005) and Fuentes and Raftery (2005). Studies on spatio-temporal data include Wikle et al. (2001); McMillan et al. (2010); Choi et al. (2009); Sahu et al. (2009, 2010); Berrocal et al. (2010b).

In air pollution science, most of the Bayesian studies on data assimilation have mainly focused on combining point-referenced data (called also geostatistical, that are data that arise from observations collected at geographical locations over a fixed continuous space) from an available network of fixed monitoring stations and output from deterministic models supplied for grid cell. Because of this misalignment in the spatial resolution of the monitoring data and the numerical model output, they faced the *change of support problem* (Gelfand et al. 2001; Gotway and Young 2002), as seen in chapter 2. In detail, the problem can be set out as in Gelfand et al. (2001). Let $Y(\mathbf{s})$ be a spatial process for locations $\mathbf{s} \in D_s$, a region of interest. The realisation of the process here could be at a finite set of sites $\{\mathbf{s}_1, \ldots \mathbf{s}_n\}$, or at block or grid cell averages. Denote by $A \subset D$ a grid cell and *Y(A)* the average exposure in a grid. Thus:

$$Y(A) = \frac{1}{|A|} \int_A Y(\mathbf{s}) d\mathbf{s}$$

where $|A|$ is the area of grid cell A. The solutions to the problem would involve different approaches.

Fuentes and Raftery (2005) proposed a melding or fusion model approach for observed data and numerical model output, performed in a purely spatial setting. The authors assumed a latent true exposure surface which was informed by both the monitoring station data and the output from the Community Multiscale Air

Quality model (CMAQ). Thus, they converted the point level $Y(\mathbf{s})$ to grid level solving the stochastic integral (i.e., integration on a average of random variables) of equation 3.1. Recently, Choi et al. (2009) extended the modelling idea of Fuentes and colleagues to incorporate a temporal dimension. The work of McMillan et al. (2010) was still framed within a fusion approach, however, instead of modelling at point level, the authors modelled at grid cell level.
A further approach was provided by Sahu et al. (2010) that proposed a model at the point rather than the grid cell level as McMillan et al. (2010). In particular, the authors formalised a latent atmospheric process which is modeled at two different scales, at the point level to align with the monitored data and at the grid cell level to align with the resolution for the deterministic model output. The models at these two scales were connected through a measurement error model. An alternative to the fusion modelling was provided by the downscaler model proposed by Berrocal et al. (2010b) that scales the output from the numerical model to point level. Specifically the authors regressed the observed point level pollution data (i.e., ozone concentrations) on the computer model output with spatially varying regression coefficients specified through a Gaussian process. The same authors extended successively the model to include the temporal dimension in a bivariate setting (Berrocal et al. 2010a).

This chapter, presents a different approach to that in the main literature on data assimilation that is best suited for combining data collected at different spatial resolutions (observed concentrations collected at point level and output from deterministic models at grid level).

Firstly, it considered a numerical model other than CMAQ, that can be output not only at grid level but also at point level, allowing for a methodology for exposure assessment working exclusively with geographically referenced data.

Secondly, it interprets the contribution of monitoring data and numerical model output as capturing different source components to the PM concentrations, and moreover considers the effect of a set of covariates with direct and indirect influ-

ence on pollution variability in urban area.

Third, it enhances the temporal dynamics of particle data collected through time, to be used in short-term health effect estimates, with the spatial features of the pollution process. In particular, space is taken to be continuous, and time is taken to be discrete and the data are viewed as PM time series at each monitoring station.

3.2 Modelling approach

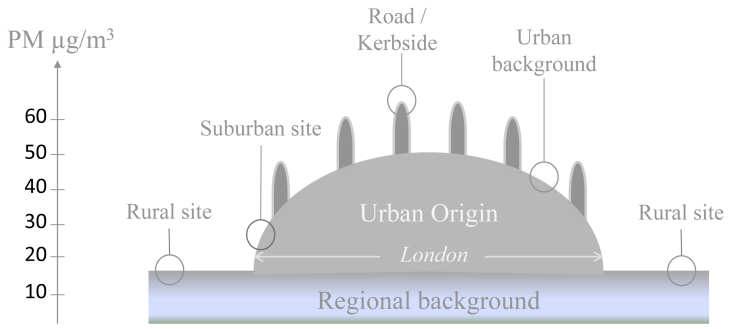
3.2.1 Environmental perspective

In the context of short-term exposure assessment, the successful modelling of PM at a local or city scale is a frequent technical challenge that requires information about regional background pollutant concentrations. This reflects the complexity of ambient PM which comprises of primary particulates arising from local traffic and non-traffic emissions, and secondary particles formed by atmospheric physical and chemical processes, such as condensation of vaporised material or by-product of the oxidation of gases, mainly during the course of long-range transport of pollutants.

From an environmental implementation perspective, the so-called Lenschow's paradigm (Lenschow et al. 2001) offered a particularly appealing approach to this problem. Lenschow and colleagues schematised the profile of PM_{10} concentrations in Berlin (Germany) using the different location's types of the monitoring sites to represent different types of exposure. This concept of different contributions (long-range transport and regional, urban and local) adapted for London is illustrated schematically as showed in Figure 3.1.

In Lenschow's study, an increasing mass concentration of PM was observed going from rural background sites to near-traffic sites (road and kerbside), passing trough suburban and urban background. PM at rural background sites could be attributed to regional and distant sources with little contribution from the agglomeration, while high PM_{10} pollution in busy streets would mostly attributed

Figure 3.1: Schematic illustration of different airborne PM_{10} concentrations in an urban area such as London (modified from Lenschow et al. (2001)).



to traffic with influence from exhaust emissions and tyre abrasion in the individual street, and resuspension of soil particles, with influence from the city background (containing also the regional background).

Therefore, particulate source is informative about the intra-urban spatial variability (Monn 2001) and, although PM represent a pollutant that is considered relatively homogeneous over space in urban environment (Shaddick and Wakefield 2002) certain particles components, especially those emitted by local traffic sources can be heterogeneously distributed (Sarnat et al. 2010).

Moreover, such separation of airborne PM by sources (such as long-range and local range components) is important as different pollution sources can give rise to particles with different chemical and physical properties which are likely to produce different effects on health outcomes. For example, Blangiardo et al. (2011) studying the association between long-range and local concentration of PM_{10} and the risk of emergency hospital admissions for cardio-respiratory diseases in Greater London during pollution episodes, found different health effect.

Following the Lenschow's paradigm, the approach proposed in this chapter considered the relative contribution of regional and local sources affecting the spatiotemporal properties of PM. Specifically:

1. A time-varying latent regional process for capturing the long-range transport of PM. The regional PM concentrations were estimated through direct measurement of rural background concentrations.

2. A spatial local process for capturing the additional urban and local primary PM component. A local scale air pollution dispersion model was used to describe this component.
3. Selected space- or time-varying factors, which could have a direct influence on the pollution process or could be used as proxies for other unmeasured variables.

3.2.2 Statistical perspective

From a statistical point of view, the implementation of models for urban PM involving variability over space and time, has been performed using a hierarchical modelling approach, which expresses the joint distribution of data, process and parameters into a series of conditional models (e.g., Wickle 2003), as discussed in chapter 2 (section 2.3.3).

Royle and Dorazio (2008) underlined that, with respect to the process model, it is possible to identify two types of hierarchical models. The first includes an *explicit* process model that describes realisations of the process; the second contains an *implicit* process model that is usually represented by random effects that are spatially and/or temporally indexed. In this chapter, both the process models for PM have been explored.

The implementation of the hierarchical approach presented here takes advantage of a space- and time-varying coefficient model. West and Harrison (1997) underlined the effective advantage of this type of approach for prediction purposes in comparison to static coefficient models.

Varying coefficient models have been proposed by Hastie and Tibshirani (1993) as a class of model where the regressor coefficients of GLMs vary as smooth functions of other variables, known as *effect modifier*. In particular, these models are characterised by an interaction effect between a smoothed function that represents a non-linear relationship and a covariate, thus generating coefficient curves (Marx 2010). In spatial statistics, these models have been described by Gelfand et al. (2003). Fan and Zhang (2008) and Park et al. (2015) presented an overview

on the major methodological and theoretical developments. Suppose $Y_1, \ldots, Y_T$ be a response variable measured at time points indexed by t, $t = 1, \ldots, T$, whose distribution depends on parameters η_t and let $\mathbf{x}_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,n})'$ and $\mathbf{u}_t = (u_{t,1}, u_{t,2}, \ldots, u_{t,n})'$ be the covariates, for $t = 1, \ldots, T$ and $i = 1, \ldots, n$. A general form of varying-coefficient model can be presented as:

$$\eta_t = b_0 + x_{t,1} b_1(u_{t,1}) + \ldots, + x_{t,n} b_n(u_{t,n}) \tag{3.1}$$

such that $u_{t,1}, u_{t,2}, \ldots, u_{t,n}$ change the coefficients (that is, they are the set of variables termed effect modifier) of $x_{t,1}, x_{t,2}, \ldots, x_{t,n}$ through the smooth and unspecified functions $b_1(\cdot), \ldots, b_n(\cdot)$.

As a special case of this general model, spatial modelling can be achieved if the effect modifiers are, for example, the geographical coordinates of locations, and a time-varying coefficient models are obtained if the effect modifier is time. In the environmental literature, there have been several applied studies using some form of spatially varying coefficients, with different statistical structures according the type of spatial data in analysis, for example Higdon (1998); Fuentes (2001) and recently Hamm et al. (2015) presented models for point-referenced data and Assunção (2003) for areal data. Regarding time-varying coefficient models, examples in time series health studies include Peng et al. (2005) that proposed seasonal models using harmonic smooth terms, Chiogna and Gaetan (2002) and Lee and Shaddick (2008) which adopted an autoregressive approach, and Lee and Shaddick (2007) that used an arbitrary smooth function.

3.3 Preliminaries on spatial point-referenced process

Before moving to the core of the chapter, several key concepts on spatial point-referenced process are briefly provided.

Point-referenced or geostatistical data are realizations of a spatial process, analogously to time series notions of a realization from a temporal process. However,

in this context space is considered as the domain rather than time. An important difference between time series and spatial data is that the latter cannot be ordered in any meaningful way. Thus, models for the spatial correlation cannot depend on the ordering of the data.

In detail, point-referenced data are realization of a spatial process or *random field*, $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ where $D \subset \mathbb{R}^d$, characterised by a spatial index $\mathbf{s}$ which varies continuously in the fixed domain D.

The mean (or expectation) of a random field is defined to be its first-order moment:

$$E[Z(\mathbf{s})] = \mu(\mathbf{s}) \tag{3.2}$$

The variance of a random field is defined as the second-order moment about the expectation $\mu(\mathbf{s})$:

$$\mathrm{Var}[Z(\mathbf{s})] = E[Z(\mathbf{s}) - \mu(\mathbf{s})] \tag{3.3}$$

A variant of the second-order moment is the covariance function that is defined as:

$$C(s_1, s_2) = E[(Z(s_1) - \mu(s_1))(Z(s_2) - \mu(s_2))] \tag{3.4}$$

for any location s_1 and s_2.

In the next sections a discussion of basic statistical concepts for characterising covariance structures of spatial processes are presented. In particular, the concept of stationarity that describes characteristics of certain (homogeneous) random fields is introduced, along with the concept of isotropy. Further, some characteristics of suitable covariance models for stationary processes are given, presenting in particular several commonly used models, notably for processes with isotropic stationary covariance structure. Methods for modeling the spatial covariance of non-stationary processes are not discussed here, as they are not relevant for this thesis, however discussions about them can be found in Le and Zidek (2006); Gelfand et al. (2010); Banerjee et al. (2014).

Stationarity

The concept of stationarity in spatial analysis is similar to that in time series analysis as seen in chapter 1 (section 1.2.2). Stationarity in simple terms means that the random field looks similar in different parts of the domain.

The spatial process is said *strongly* (or *strictly*) *stationary* if for any given set of locations, $\{\mathbf{s}_1, \ldots \mathbf{s}_n\}$, and any displacement, $\mathbf{h} \in \mathbb{R}^d$, the distribution of $(Z(\mathbf{s}_1), \ldots, (Z(\mathbf{s}_n))$ is the same as $(Z(\mathbf{s}_1 + \mathbf{h}), \ldots, (Z(\mathbf{s}_n + \mathbf{h}))$.

A less restrictive concept is given by the connotation of the process as *weakly stationary* (or *second order stationary*). In particular, a spatial process is weakly stationary if:

- $E(Z(\mathbf{s})) = E[Z(\mathbf{s} + \mathbf{h})] = \mu$
- $\text{Cov}(Z(\mathbf{s}), Z(\mathbf{s} + \mathbf{h})) = C(\mathbf{h})$

for every $\mathbf{h} \in \mathbb{R}^d$. Thus, this type of spatial process has constant mean and covariance that depends only upon the displacement vector $\mathbf{h}$. This assumption means that the variability of the spatial process is the same everywhere. The graphical representation of covariance function, $C(\mathbf{h})$, is sometimes called a *covariogram*.

At $\mathbf{h}$ =0, it would be $\text{Cov}(Z(\mathbf{s} + 0), Z(\mathbf{s})) = C(0) = \text{Var}(Z(\mathbf{s}))$. Hence, the weakly stationarity implies the existence of the variance that does not depend on the location (Le and Zidek 2006).

The strong stationarity implies weak stationarity, but the reverse is not implied.

Finally, a spatial process is said *intrinsic stationary* if the mean is constant and the difference $(Z(\mathbf{s} + \mathbf{h}) - Z\mathbf{s}))$ is second order stationary. Thus $E[Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})] = 0$, then the $\text{Var}[Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})]$ depends on only the sites relative positions $(\mathbf{s} + \mathbf{h}) - \mathbf{s}$, i.e., there is a function γ such that: $\text{Var}[Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})] = 2\gamma[(\mathbf{s} + \mathbf{h}) - \mathbf{s}] = 2\gamma(\mathbf{h})$.

The function $2\gamma(\mathbf{h})$ is called *variogram* and $\gamma(\mathbf{h})$ is called *semivariogram*. The second order stationarity implies intrinsic stationarity but the reverse is not implied. There is a similarity between intrinsic stationarity and second order stationarity, such that the intrinsic is defined in terms of the variogram and second

order is defined in terms of the covariance function. In particular, the variogram is a generalization of the covariance function and under second order stationarity the two functions are related (e.g., Cressie and Wikle 2011):

$$\begin{aligned}
\gamma(\mathbf{h}) &= \frac{1}{2}E[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2] \\
&= \frac{1}{2}E[((Z(\mathbf{s}+\mathbf{h}) - \mu) - (Z(\mathbf{s}) - \mu))^2] \\
&= -E[(Z(\mathbf{s}+\mathbf{h}) - \mu)(Z(\mathbf{s}) - \mu)] + \frac{1}{2}E[(Z(\mathbf{s}+\mathbf{h}) - \mu)^2] + \frac{1}{2}E[(Z(\mathbf{s}) - \mu)^2] \\
&= -C(\mathbf{h}) + C(\mathbf{0}) \\
&= C(\mathbf{0}) - C(\mathbf{h})
\end{aligned}$$

Thus, the variogram describes the degree to which nearby locations have similar values. The plot of the semivariance, that is formally defined as the squared difference in height between locations, $\hat{\gamma}(\mathbf{h}) = \frac{1}{2n(\mathbf{h})}\sum_{i=1}^{n}(Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h}))^2$ versus the lag distance is the empirical variogram, which is subsequently modeled by a parametric model. Common examples of isotropic variogram models include the linear, the spherical, the exponential, the powered exponential, the Matérn (e.g., Banerjee et al. 2014). The characteristic parameters of the variogram are the *sill*, that is the semivariance value at the plateau of the variogram, representing the distance in which the observations are no longer correlated, the *range*, that is the lag where the semivariance is equal to the sill and the *nugget*, that represents the minimum variance and is the semivariance value at lag distance equal to zero. (Deligiorgi and Philippopoulos 2011).

Isotropy

The concept of *isotropy* in space is defined as invariance under rotation about a given spatial location (Cressie and Wikle 2011). In detail, a second-order stationary spatial process is said isotropic if its covariance function depends upon the separation vector only through the distance $C(\mathbf{h}) = C(\|\mathbf{h}\|)$ for all $\mathbf{h}$, where $\|\cdot\|$ indicate the Euclidean distance. This assumption implies that the covariance between observations located at $\|\mathbf{h}\|$ units apart is the same, independently of the location and geographical direction (north-south or east-west).

When covariance functions exhibit different behavior in different directions, the random fields are called *anisotropic*.

This issue of anisotropy does not have equivalence in time series (Schabenberge and Gotway 2004).

Covariance function

The necessary conditions for C to be a covariance function are (Le and Zidek 2006):

- $C(\mathbf{0}) \geq 0$, since $C(\mathbf{0}) = \text{Var}[Z(\mathbf{s})] \geq 0$;
- $C(\mathbf{h}) = C(-\mathbf{h})$ for any vector $\mathbf{h}$ since the covariance is an even function;
- $C(\mathbf{0}) \geq |C(\mathbf{h})|$, where $|\cdot|$ denotes the absolute value (this inequality derives by applying Schwarz's inequality (Shorack and Wellner 1986)).

Furthermore, a key property satisfied by the covariance function is that nonnegative definiteness, as seen in chapter 1 for the time series process. In detail, for any a_i and $a_j \in \mathbb{R}$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(s_i - s_j) \geq 0$$

The nonnegative definiteness condition is necessary for the existence of a random field with finite second moments. This condition guarantees that the variance of spatial predictions is non-negative.

There are a variety of approaches for modelling spatial covariance structures. Banerjee et al. (2014) discuss several popular examples of isotropic covariance functions. Among them, the Matérn family of covariance functions (Matérn 1986) provides a very general choice, allowing control of spatial association. Denote for notation simplicity $\|\mathbf{h}\|$ by d. The Matérn covariance function between locations is given by:

$$C(d) = \frac{\omega^2}{2^{\nu-1}\Gamma(\nu)}(\phi d)^{\nu} B_{\nu}(\phi d),\ \phi > 0, \nu > 0, d > 0 \tag{3.5}$$

where ω^2 is the marginal variance, ν is the *smoothness parameter* controlling the smoothness of the process (where, higher values yield smoother process realiza-

tions), B_ν is the modified Bessel function of order ν, ϕ is the *decay parameter* that controls the range of spatial correlation and Γ is the Gamma function. The smoothness of a random field, the parameter ν in the Matérn class, plays a critical role in interpolation problems. A number of commonly used models for the covariance structure, including Gaussian and exponential structures assume that the smoothness parameter is known a priori. The Gaussian model is the limiting case of the Matérn model as $\nu \to \infty$. The exponential model, that will be described in more detail in section 3.6 is a special case of Matérn model with smoothness parameter fixed at ν =0.5.

Spatial models and predictions

In geostatistics often the primary interest is given by a prediction problem. Predicting unmeasured responses at locations of interest, using observations made at sites over the geographical domain in study, is commonly called *spatial interpolation* or *spatial prediction*. The aim here is to predict $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ for $D \subseteq \mathbb{R}^2$ at new location $\mathbf{s}_0 \notin \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$.

There are several techniques for spatial interpolation of geostatistical data. *Kriging* is by far one of the most commonly used. It defines a collection of methods for spatial interpolation. The term was coined by Matheron (1962) in honour of the South African engineer D. G. Krige who inspired the general approach for mining applications. In the original formulation of Kriging no distributional assumptions were performed. Models for spatial data were introduced by Matheron (1962) and successively they were popularised by Cressie (1993). In the classical approach, the *Gaussian process* represents the common framework.

Gaussian process is a type of stochastic process that is well-suited for spatial and spatio-temporal modelling. It is a continuously defined process such that all the finite-dimensional distributions are multivariate Gaussian (or Normal) distributions (Rasmussen and Williams 2006). It can be viewed as infinite-dimensional Gaussian distributions. While a Gaussian distribution is a distribution over scalars or vectors (for multivariate distributions), and it is fully specified by a mean

and a covariance matrix, a Gaussian process is a distribution over functions and it is uniquely defined by a mean function and a covariance function. It has been largely used as method for nonparametric regression and classification tasks (Rasmussen and Williams 2006).

Within the field of geostatistics, a spatial process $\{Y(\mathbf{s})\}$ is formally defined as a Gaussian process if the joint distribution of $(Z(\mathbf{s}_1), \ldots, (Z(\mathbf{s}_n))$ is multivariate Normal distribution for any finite subset $\{\mathbf{s}_1, \ldots \mathbf{s}_k\}$ of locations $\mathbf{s}$. For a Gaussian process, the concepts of strong and weak stationarity previously described are equivalent.

Indeed, the Kriging methods assumes a Gaussian process structure for the unknown spatial field and focuses on calculating the optimal linear predictor of the field. It is essentially a weighted linear combination of observed values, where neighboring sites are assigned weights such that the prediction error is minimized. A general linear model for the random field $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is given by:

$$Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \epsilon(\mathbf{s}_i) \tag{3.6}$$

where $\mu(\mathbf{s}_i)$ is a deterministic function and $\epsilon(\mathbf{s}_i)$ is Normally distributed with mean zero and covariance matrix, $\text{diag}(\Sigma) = \{\sigma^2, \ldots, \sigma^2\}$.

Kriging is an optimal linear estimator of $Z(\mathbf{s}_0)$ that takes the form:

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \alpha_i Z(\mathbf{s}_i) \tag{3.7}$$

where $\boldsymbol{\alpha} = (\alpha_1(\mathbf{s}), \ldots, \alpha_n(\mathbf{s}))$ is the vector of Kriging weights that are computed such that $\hat{Z}(\mathbf{s}_0)$ is regarded as the best linear unbiased predictor, in the sense that its error has minimal variance among all linear combinations of the observations. The weights are obtained as:

$$\boldsymbol{\alpha} = \mathbf{C}^{-1}\rho \tag{3.8}$$

where $\rho = [\text{Cov}(Z(\mathbf{s}_0), Z(\mathbf{s}_1), \ldots, \text{Cov}(Z(\mathbf{s}_0), Z(\mathbf{s}_n))]'$ and $\mathbf{C}^{-1} = \text{Cov}\{[Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)]'\}$.

Kriging appears in many forms. The most common Kriging methods are the *simple Kriging*, which assumes a known constant mean, *ordinary Kriging*, which

assumes an unknown constant mean, estimated from the data and *universal Kriging* which assumes an unknown mean that is a function of $\mathbf{s}$.

In all the cases, the classical implementation of the Kriging includes two stages: the analysis of the spatial variation performed through the variogram assessment and the prediction of the target variable, plugging in the estimated variogram parameters into the Kriging equation (Deligiorgi and Philippopoulos 2011).

In classical Kriging, the aspect of the uncertainly associated with the model parameters is considered only in a marginal way, ignoring it in the subsequent predictions. The Bayesian approach assumes, instead, that the parameters are unknown and treats these as random variables and integrates over the parameter space to obtain the predictive distribution of any quantity of interest (e.g., Le and Zidek 2006; Banerjee et al. 2014). Following the approach to interpolation problem as in Banerjee et al. (2014), the Gaussian spatial process can be written as a GLM:

$$Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + u(\mathbf{s}_i) + \epsilon(\mathbf{s}_i) \tag{3.9}$$

where $\mu(\mathbf{s}_i)$ is the mean function at location $\mathbf{s}_i$, $i = 1, \ldots, n$. The residual $Z(\mathbf{s}_i) - \mu(\mathbf{s}_i)$ is partitioned in two components: the correlated error term, $u(\mathbf{s}_i)$, called the partial sill and the uncorrelated error term, $\epsilon(\mathbf{s}_i)$, called the nugget effect (it represents measurement error and/or microscale variability). The vector $\mathbf{u} = (u(s_1), \ldots, u(s_n))$ is assumed to be Normally distributed with mean zero and covariance matrix Σ equal to ω^2H, where H is a correlation matrix that takes into account for the spatial correlation. The error term is Normally distributed with mean zero and variance σ^2_ϵ.

Bayesian Kriging can be achieved assigning a prior probability distributions to the unknown quantities of the model. Then, let θ be a vector of all model parameters. The posterior predictive distribution of the observation at an unobserved site $\mathbf{s}_0$ is given by:

$$p(z(\mathbf{s}_0)|\mathbf{z}) = \int p(z(\mathbf{s}_0)|u(\mathbf{s}_0), \theta) p(u(\mathbf{s}_0)|\mathbf{u}, \theta) p(\mathbf{u}, \theta|\mathbf{z}) du(\mathbf{s}_0) d\mathbf{u} d\theta \tag{3.10}$$

3.4 Description of the data

The PM_{10} data ($\mu g/m^3$) available for this study were daily average concentrations (midnight to midnight) collected in the years 2002-2003 (728 days). This period was selected to include several winter and summer pollution episodes (defined as periods where the background air pollution is unusually high for a sustained period) and also the 2003 European heat-wave (Johnson et al. 2005; Solberg et al. 2005). They came from three sources:

1. *Mass concentration measurements from the London Air Quality Network (LAQN).* This monitoring network had 76 PM_{10} sites in 2002-2003, with some of these also affiliated with the National Automatic Urban and Rural Network (AURN). Out of these sites, we selected 45 for which the proportion of missing data, in each year, did not exceed 20%. The missing observations were assumed to be missing at random. The average proportion of missing data for the 45 sites in the study period was 5.1 % (range: 0-17.4).

   The majority of measurements were made using the Tapered Element Oscillating Micro-balance method using TEOM 1400a and 1400ab analysers (R&P Tapared Element Oscillating Microbalance). These instruments are known to underestimate the PM_{10} concentrations due to losses of semi-volatile constituents (such as ammonium nitrate and organic aerosols) (Allen and Reiss 1997; Green et al. 2001) therefore measurements from TEOM analysers were multiplied by a conversion factor of 1.3 (DETR 1999). A dynamic correction has been available since 2004 using measured concentrations of volatile PM_{10} (Green et al. 2009). Eight sites were equipped with Beta Attenuation Monitors (Met-One BAM) where a correction factor of 0.82 was applied according to the results of UK trails (Harrison 2006) that compared the Met-One BAM to a reference instrument.

2. *Output from the high spatial resolution Air Dispersion Modelling System*

(ADMS-Urban; CERC, Cambridge, UK)[1] (McHugh et al. 1997; Carruthers et al. 2000). ADMS-Urban was used to represent the local primary component of PM_{10}. It simulates the dispersion into the atmosphere of pollutants released from road traffic, industrial and domestic sources across urban areas and integrates emissions inventories with meteorological data. Emissions factors were obtained from the London Atmospheric Emissions Inventory, which contains data on road network geometry comprising about 60,000 individual road links attributed with traffic flows and composition. Roads are represented as line sources in ADMS-Urban with a spatial precision of less than one metre. Point and area source emissions were aggregated in the London Atmospheric Emissions Inventory to one kilometre resolution grids. This is a relatively quick method for modelling poorly defined or diffuse sources in the dispersion model (e.g., domestic heating). Dispersion from road sources used a Gaussian plume model with a non-Gaussian plume profile in convective conditions to account for the skewed structure of the vertical component of turbulence. Grid sources were modelled using a simple trajectory model. Output from both models was combined to predict pollutant concentrations at point locations, namely air pollution monitoring sites. Meteorological data (wind speed, wind direction, temperature, and cloud cover) included in ADMS-Urban were obtained from the British Atmospheric Data Centre for the nearest site.

Because ADMS-Urban is fully integrated with GIS, it allowed for spatial point estimates. This feature avoided the change of support problem.

3. *Mass measurements from rural monitoring sites.* Background concentrations, as proxy of the long-range transport of PM_{10}, were sourced from two rural monitoring stations belonging to the AURN, approximately equidistant from London: (i) Harwell (near Didcot, Oxfordshire), 81 km north-west of central London, towards the West Midland conurbation; and (ii) Detling

[1]The ADMS-Urban output used in the modelling approach described in this Chapter was provided by John Gulliver of Imperial College London, UK.

(Kent), 50 km south-east of central London towards continental pollution sources areas. The sites were chosen to provide different information about the long-range transported air pollution affecting London.

Additionally, the following set of covariates was available for the analysis:

1. *Type of site* which accounted for different environmental conditions. The LAQN monitoring sites were classified into different types, depending on their location. Of the 45 sites selected for the study, eight were suburban sites (located in residential areas on the outskirts of London), 13 were urban background sites (located away from major sources and broadly representative of city-wide background concentrations), 20 were roadside sites (located from one and five metres from a major carriageway) and four were kerbside sites (located within one metre of a major road carriageway).

2. *Day of the week* which accounted for unknown changes in emissions between weekdays and weekend days, because emission inventories are not time-varying but only contain annual totals. The indicator variable for day of the week was categorised as Monday-Friday, Saturday, and Sunday or Public Holiday.

3. *Average daily temperature* to describe seasonal changes in chemistry between primary and regional secondary PM_{10}. Other meteorological variables were not considered since these are used in the ADMS-Urban model, however this does not include secondary PM_{10} formation, hence daily mean temperature was used as a surrogate for such processes. Over the 2002-2003 years, the average temperature, recorded at London Heathrow, was 11.9°C , with daily mean ranging between -1.3°C and 28.2°C.

3.4.1 Data processing

Airborne particle measurements from the LAQN and the AURN, as well as ADMS-Urban output were transformed using the logarithmic scale, since their

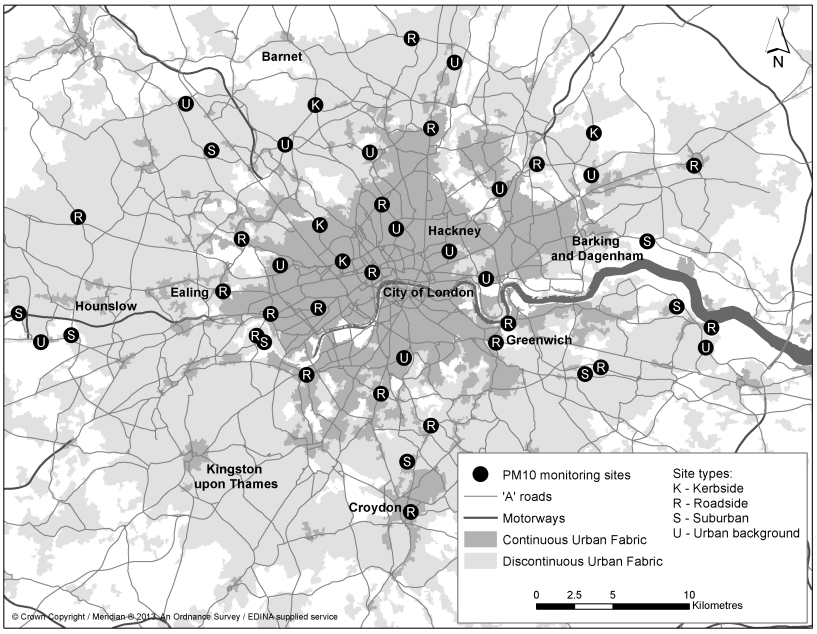
distributions tend to be positively skewed. There is some justification in supporting this transformation, given by physical explanation of atmospheric chemistry (Ott 1990). In particular, after the pollutants are emitted from the source(s), in the transport process before they reach the receptor site, they undergo successive mixing and diluting, that result in a log-normal distribution (Kan and Chen 2004).

3.5 Exploratory data analysis

Figure 3.2 presents the geographical location of the monitoring sites across Greater London by site type. This monitoring networks showed an irregular design as monitors are heavily concentrated in the city centre, with less dense coverage in the surrounding areas.

Because little difference were found between the PM_{10} concentrations at suburban and urban background sites, these two categories were aggregated.

Figure 3.2: Location and siting characteristics of the air quality monitoring sites in Greater London selected for the study.



The mean distance between the selected sites was 16,967 metres (range: 657 - 45,298)[2]. Figure 3.3 shows the correlation of daily data for pairs of monitoring

[2]The distance matrix was computed using the haversine formula (as implemented in the R

sites as a function of their separation distance. The correlations were generally high, also over long distances ($\geq$ 30,000 m), indicating that factors other than distance may have a role in explaining the spatial variability of PM_{10} concentrations.

Figure 3.3: Correlation between pairs of monitoring sites as a function of their separation distance.

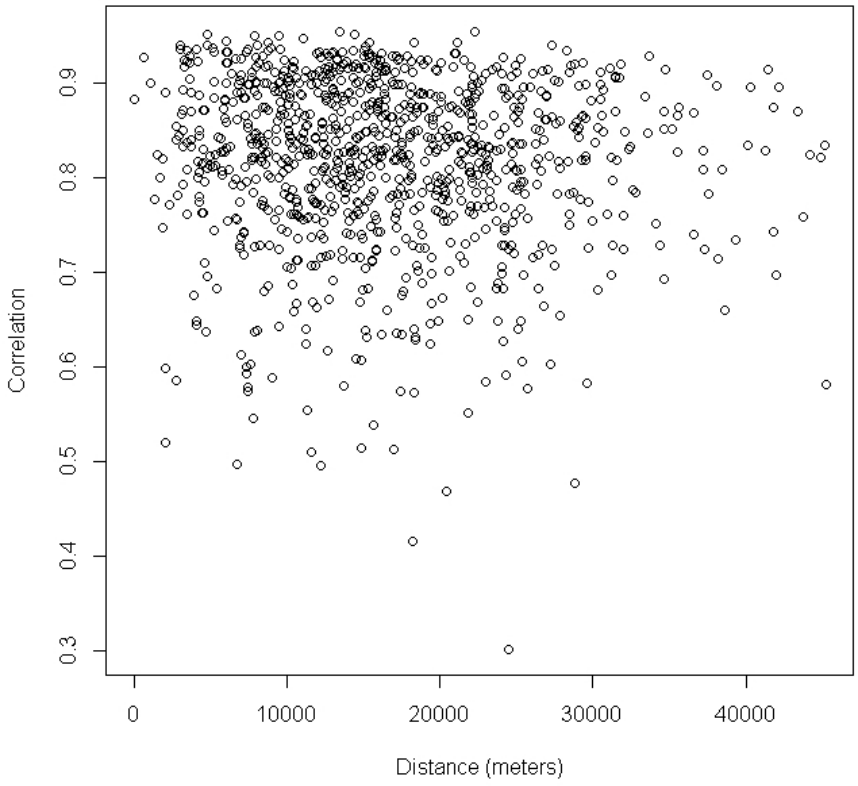


Figure 3.4 presents the daily concentrations of PM_{10} across the 45 monitoring sites sorted from the top to the bottom by decreasing longitude (from west to east), during the two year study period. The daily values are displayed according to the tertiles computed on the global data set to ensure the comparability of the

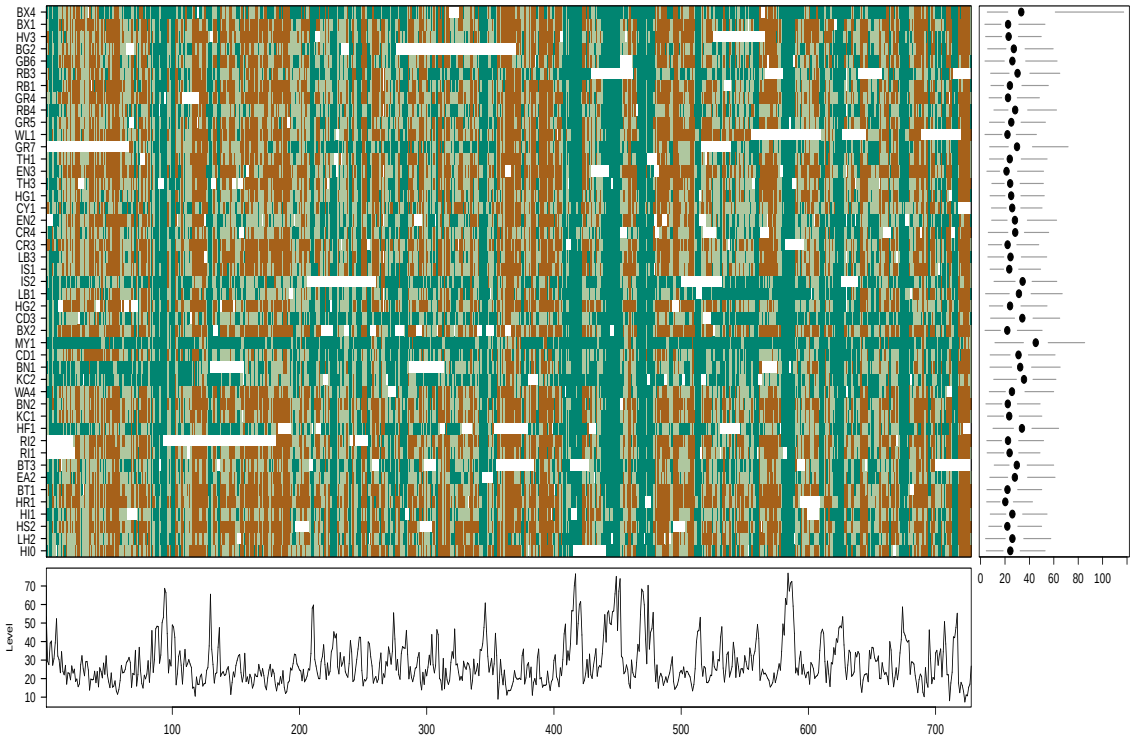
package *geosphere*), that mathematically is as follows:

$$a = \sin^2(\frac{\Delta\varphi}{2}) + \cos\varphi_1 * \cos\varphi_2 * \sin^2(\frac{\Delta\lambda}{2})$$
$$c = 2 * \mathrm{atan2}(\sqrt{a}, \sqrt{(1-a)})$$
$$d = R * c$$

where φ is the latitude, λ the longitude, Δ indicates their absolute difference and R is the radius of the earth (mean radius = 6,371km); the angles are in radians.

time series and assigned to low (brown), medium (pale green) and high (green) categories of PM_{10} concentrations (Peng 2008). Missing data are denoted by the colour white. The bottom of the plot shows the daily median values across all the time series and the right hand side panel the boxplots of the data in each time series.

Figure 3.4: Daily particle concentrations for the 45 monitoring sites sorted from the top to the bottom by decreasing longitude.

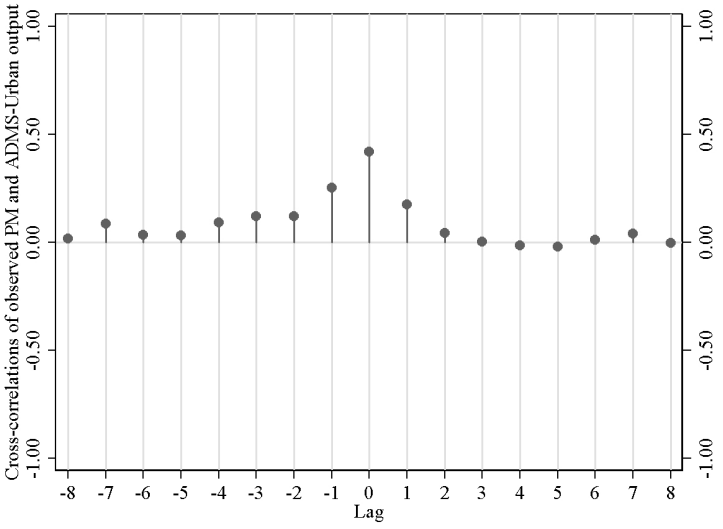


The PM_{10} pollution episodes that London experienced during February, March, April, August, September and November 2003 are clearly visible. These episodes were mainly caused by secondary PM_{10} from distant sources, with summer episodes also being linked to photochemistry (Fuller and Green 2006). The November 2003 episode was associated with Guy Fawkes Night fireworks and bonfires (Fuller 2003).

The analysis via cross-correlogram of the time series of PM_{10} concentrations observed in Greater London and the local component of PM_{10} captured from ADMS-Urban output, presented in Figure 3.5, shows that the correlation was relatively high and positive at lag 0 (same day pollution concentrations), sug-

gesting that the numerical model captured the time variation of PM_{10} observed at monitoring sites.

Figure 3.5: Cross-correlogram between the time series of particle concentrations in Greater London and the ADMS-Urban output (on log-scale).



Finally, a graphical check of the relationship between PM_{10} concentrations and temperature was performed through a scatterplot which showed a nonlinear relationship.

3.6 Model specification

Denote $\mathbf{y}_{\mathbf{s},t}$ be the log-transformed daily PM_{10} concentrations, with $\mathbf{s} = 1, \ldots, n$ =45 (sites of the pollutant monitoring network) and $t = 1, \ldots, T$ =728 (days). The data model consisted in a Gaussian measurement error, that is:

$$\mathbf{y}_{\mathbf{s},t} = \mu_{\mathbf{s},t} + \epsilon_{\mathbf{s},t} \tag{3.11}$$

where $\mu_{\mathbf{s},t}$ represents the mean process driven by covariates varying over space and time and $\epsilon_{\mathbf{s},t}$ are site-specific zero-centred Gaussian disturbances, such that: $\epsilon_{\mathbf{s},t} \sim \mathrm{N}(0, \sigma^2_{\mathbf{s}})$ (Shaddick and Wakefield 2002; Cocchi et al. 2007).

A class of different nested statistical formulations for the mean space-time process, $\mu_{\mathbf{s},t}$, was considered. These structures differently accounted for factors affecting the spatio-temporal properties of particle concentrations.

This first model represented a simple statistical structure where the daily measurements at each monitoring site were assumed to be a function of a residual mean concentration across the urban area and a latent pollutant process described by the long-range transported component of particulate. The time-varying latent regional process was included assuming that concentrations at the city scale derive largely from information borrowed from background measurements. It assumed the form:

$$\textit{Model 1} : \mu_{\mathbf{s},t} = \alpha + \mu_t^{lrt} \tag{3.12}$$

where α is the residual intercept and μ_t^{lrt} represents the mean of the latent process.

In particular, let $\mathbf{j}$ denote several available rural background monitoring sites around the metropolitan area, with $\mathbf{j} = 1, \ldots, J$. The model assumed that the time series of pollution data from the rural monitoring sites were a reflection of an underlying long-range transport of particles into the urban area, measured with error:

$$lrt_{\mathbf{j},t} \sim \mathrm{N}(\mu_t^{lrt}, \sigma^2_{lrt,\mathbf{j}}) \tag{3.13}$$

In this application, this latent process was driven by the concentrations of PM_{10} measured at the Harwell and Detling rural background sites ($\mathbf{j}$ =1,2).

This simple model accounted for the temporal variability of the pollution process, but did not incorporate a spatial structure. The model describes the main hypothesis in the definition of air pollution exposure in ecological time series studies, where the pollution estimates for a given study region, are generally free from a spatial dimension, although these studies typically use averaged ambient pollutant levels from one or more background monitoring stations to represent the exposure experienced by a study population.

The second model considered, added to the constant, α, the local city primary

PM_{10} component described by ADMS-Urban, $\ell_{\mathbf{s},t}$:

$$Model\ 2: \mu_{\mathbf{s},t} = \alpha + \beta_{1,\mathbf{s}}\ell_{\mathbf{s},t} \tag{3.14}$$

To capture the spatially varying effects of the local primary component of PM, a spatially varying coefficients model was assumed. In this way, the model allowed the regression parameters for the ADMS-Urban, $\boldsymbol{\beta}_1 = (\beta_{1,1}, \ldots, \beta_{1,n})'$, to be different in different sites trough a varying slope implemented using a Gaussian isotropic kriging model (Banerjee et al. 2014). In particular, the vector parameter $\boldsymbol{\beta}_1$ was specified as a zero-mean multivariate Normal distribution

$$\boldsymbol{\beta}_1 \sim MVN(\mathbf{0}, \Sigma) \tag{3.15}$$

where, $\mathbf{0}$ is a zero vector and Σ is the positive definite spatial covariance matrix specified as follows:

$$\Sigma = \omega^2 \mathrm{H}(\phi), \quad \phi > 0 \tag{3.16}$$

Here, ω^2 is the spatial effect variance parameter, H is the nxn spatial correlation matrix and ϕ is the decay parameter. The correlation between $\beta_{1,\mathbf{s}_i}$ and $\beta_{1,\mathbf{s}_j}$ is a function of their geographic separation, that can be specified in several ways, as described in section 3.3. The correlation functions are described by a exponential function:

$$\mathrm{H}_{ij}(\phi) = \exp(-\phi d_{ij}) \tag{3.17}$$

where d_{ij} is the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, that is, $d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$.

The assumptions of stationarity and isotropy, that is constant mean, and that the relationship between the measured PM_{10} and the output from ADMS-Urban is not varying over the geographic domain, and the correlation function is supplied as a function of the separation distance between sites could be restrictive. However, this assumption looks realistic for an urban environment as Greater London, since the meteorology and topography are relatively spatially stable (Shaddick and Wakefield 2002).

The third model included both the regional and the local primary PM_{10} components:

$$\textit{Model 3}: \mu_{\mathbf{s},t} = \alpha + \mu_t^{lrt} + \beta_{1,\mathbf{s}}\ell_{\mathbf{s},t} \tag{3.18}$$

Both these components were described with the structures specified for the previous two models.

Model 4 was performed to explore the effect of the set of covariates (without regional and local PM_{10} components):

$$\textit{Model 4}: \mu_{\mathbf{s},t} = \alpha + \beta_{2,\text{type}_\mathbf{s}} + \beta_{3,\text{dow}_t} + \beta_{4,t}\text{temp}_t \tag{3.19}$$

where "type" is the type of site, "dow" is the day of the week, and "temp" is the temperature. In particular, site type was used to represent possible difference in concentration levels, as road and kerb sites are likely to have higher concentrations as they are closer to traffic source of pollution; daily mean temperature to describe chemical processes affecting local PM_{10} concentrations which are not considered in local-scale dispersion models and day of week to account for time varying emission rates which are not described in emissions inventories.

In (3.19) the fixed effects coefficients β_2 and β_3 are unknown parameters for the variables site type and day of the week. The vector $\boldsymbol{\beta_4} = (\beta_{4,1}, \ldots, \beta_{4,T})'$ is the dynamic parameter associated with the temperature, stochastically built according to a Gaussian second order random walk (RW2), which was found provide the best smoothness prior for this variable. Thus the general form of the prior would be:

$$\beta_{4,t} \sim \text{N}(2\beta_{4,t-1} - \beta_{4,t-2}, \sigma_v^2) \tag{3.20}$$

In practice, for the present study, the non-stationary RW2 model was represented using an intrinsic Gaussian conditional autoregressive prior (Fahrmeir and

Lang 2001), which produces centred effects:

$$\beta_{4,t}|\beta_{4,-t} \sim \begin{cases} \mathrm{N}\left(2\beta_{4,t+1} - \beta_{4,t+2}, \sigma_v^2\right) & \text{for } t = 1 \\ \mathrm{N}\left(\frac{2\beta_{4,t-1}+4\beta_{4,t+1}-\beta_{4,t+2}}{5}, \frac{\sigma_v^2}{5}\right) & \text{for } t = 2 \\ \mathrm{N}\left(\frac{-\beta_{4,t-2}+4\beta_{4,t-1}+4\beta_{4,t+1}-\beta_{4,t+2}}{6}, \frac{\sigma_v^2}{6}\right) & \text{for } t = 3, \ldots, T-2 \\ \mathrm{N}\left(\frac{-\beta_{4,t-2}+4\beta_{4,t-1}+2\beta_{4,t+1}}{5}, \frac{\sigma_v^2}{5}\right) & \text{for } t = T-1 \\ \mathrm{N}\left(-\beta_{4,t-2} + 2\beta_{4,t-1}, \sigma_v^2\right) & \text{for } t = T \end{cases} \tag{3.21}$$

where $\beta_{4,-t}$ represents the vector of β_4's with $\beta_{4,t}$ removed and σ_v^2 is the conditional variance. A non-stationary RW2 acts as a smoothness prior based on the second difference and penalises deviations from a linear trend (Lee and Shaddick 2008). This prior, for regular time-point, provides enough flexibility due to its invariance under addition of a linear trend and it is computationally convenient due to its Markov properties (Lindgren and Rue 2008). The choice of this prior followed also the initial explorative analysis, where we found that the relationship between temperature and PM_{10} concentrations, was potentially well described by a cubic smoothing spline. The RW2 is a discrete-time analogue of a cubic smoothing spline (Chiogna and Gaetan 2002). In the sensitivity analysis, the RW2 was assessed in term of performance in comparison to a thin-plate smoothing spline.

Model 5, finally represented the full model that accounted for the regional and local PM_{10} components and for the covariates:

$$\textit{Model 5}: \mu_{\mathbf{s},t} = \alpha + \mu_t^{lrt} + \beta_{1,\mathbf{s}}\ell_{\mathbf{s},t} + \beta_{2,\mathrm{type}_{\mathbf{s}}} + \beta_{3,\mathrm{dow}_t} + \beta_{4,t}\mathrm{temp}_t \tag{3.22}$$

Other parameter priors and hyperpriors

A Gaussian prior distribution with mean zero and variance 10^2 was assigned to the intercept α, and to the fixed effects coefficients β_2 and β_3. To ensure identifiability, we fixed the first category of these two parameters as zero ($\beta_{2,1} = 0$ and $\beta_{3,1} = 0$). The same Gaussian prior was chosen for the mean of the latent background process. Weakly informative inverse-Gamma (IG) hierarchical priors

were specified for the error variance parameters $\sigma^2_{\mathbf{s}} \sim \text{IG}(a, b)$, $\mathbf{s} = (1, \ldots, n)$, $\sigma^2_{lrt_{\mathbf{j}}} \sim \text{IG}(c, d)$, $\mathbf{j} = (1, \ldots, J)$, setting the hyperpriors ($a = c, b = d$) as IG(1, 0.1).

Similarly, inverse-Gamma priors were specified for the between-site variance component, ω^2, and the variance of the RW2, σ^2_v, with hyperparameters IG(1, 0.1).

A discrete uniform prior distribution was assumed for ϕ, the decay parameter in the spatial correlation, as suggested by Diggle and Ribeiro (2007) with range chosen based on prior beliefs about the minimum and maximum correlation at the smallest and largest distances. Typically, locations close in space are assumed to be characterised by a stronger degree of correlation, but a strong prior was not assumed, allowing for a range of correlation between 0.10 and 0.99. For large separation distances a range between 0.01 and 0.65 was specified.

3.6.1 Comparison with models implemented with varying intercepts

The model formulation proposed here deviates from the standard spatio-temporal statistical models that include varying intercepts (baseline concentrations) that are spatially or temporally correlated (Gelman and Hill 2007). The most common setting (e.g., Shaddick and Wakefield 2002; Sahu et al. 2006; Cocchi et al. 2007; Berrocal et al. 2010b) assumes that the spatial and temporal dependencies are introduced into the modelling in the form of random effects. Thus, pollution concentrations characterised by a Gaussian likelihood, are typically related to a trend surface model together with additive independent random spatiotemporal effects that in a simple implementation can assume the form:

$$\mu_{\mathbf{s},t} = \boldsymbol{\beta}\mathbf{x}_{\mathbf{s},t} + \theta_t + \eta_{\mathbf{s}} + \epsilon_{\mathbf{s},t} \tag{3.23}$$

Here, $\boldsymbol{\beta}$ is a vector of regression coefficients associated with the covariates $\mathbf{x}(\mathbf{s}, t)$. The residual is partitioned into a temporal, θ_t, a spatial, $\eta_{\mathbf{s}}$, and an independent process $\epsilon_{\mathbf{s},t}$ which is Gaussian with zero-mean and $\upsilon^2_{\mathbf{s}}$ variance. As a comparison with the proposed approach, a model implementation within this classical

framework using the same set of data has been considered. Five nested hierarchical structures that incorporated separable random space and time effects were developed as follow.

The first model includes only the spatial and temporal intercepts as random effects:

$$Model\ I: \mu_{\mathbf{s},t} = \theta_t + \eta_{\mathbf{s}} \tag{3.24}$$

The parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)'$ should capture the residual temporal dynamics characterising the pollutant process. This temporal process was described using a first-order non-stationary random walk model as daily dependence on air particulate concentrations can be expected (Shaddick and Wakefield 2002), and was built as:

$$\theta_t|\theta_{-t} \sim \begin{cases} \mathrm{N}\left(\theta_{t+1}, \sigma_\theta^2\right) & \text{for } t = 1 \\ \mathrm{N}\left(\frac{\theta_{t-1}+\theta_{t+1}}{2}, \frac{\sigma_\theta^2}{2}\right) & \text{for } t = 2, \ldots, T-1 \\ \mathrm{N}\left(\theta_{t-1}, \sigma_\theta^2\right) & \text{for } t = T \end{cases} \tag{3.25}$$

where θ_{-t} all elements of θ with θ_t removed and σ_θ^2 is the conditional variance. The term $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ represents a spatially varying intercept described by a zero-centered Gaussian process with variance σ_η^2 and exponential correlation function that depend upon the inter-site distance and the parameter ϕ quantifying the correlation decay.

Model II included also μ_t^{lrt}, as defined in equation 3.13, that is the latent regional process capturing the long-range transport component of PM_{10}:

$$Model\ II: \mu_{\mathbf{s},t} = \theta_t + \eta_{\mathbf{s}} + \mu_t^{lrt} \tag{3.26}$$

Model III added to the random effects the urban local component of PM_{10} described by ADMS-Urban:

$$Model\ III: \mu_{\mathbf{s},t} = \theta_t + \eta_{\mathbf{s}} + \beta_{1,\mathbf{s}}\ell_{\mathbf{s},t} \tag{3.27}$$

The space-varying slope $\boldsymbol{\beta_1} = (\beta_{1,1}, \ldots, \beta_{1,n})$ was build according to a Bayesian kriging (Banerjee et al. 2014) as specified in the main analysis.

Model IV incorporated both the long-range and the local components of PM_{10}:

$$Model\ IV: \mu_{\mathbf{s},t} = \theta_t + \eta_{\mathbf{s}} + \mu_t^{lrt} + \beta_{1,\mathbf{s}}\ell_{\mathbf{s},t} \tag{3.28}$$

Model V included exclusively the spatio-temporal random intercepts and the covariates type site, day of the week and daily mean temperature:

$$Model\ V: \mu_{\mathbf{s},t} = \theta_t + \eta_{\mathbf{s}} + \beta_{2,\mathrm{type}_{\mathbf{s}}} + \beta_{3,\mathrm{dow}_t} + \beta_{4,t}\mathrm{temp}_t \tag{3.29}$$

Similarly to the main analysis, a full model including the set of covariates as well as the long-range transport and local primary components of PM_{10} was also implemented, however it resulted over-parameterised and yielded implausible predictions (that is, some negative predictions).

Models I-V were specified assuming for the variance parameters σ_θ^2 and σ_η^2 inverse-Gamma priors, IG $\sim (1, 0.1)$. The remaining priors were specified as in the main analysis.

3.6.2 Performance assessment

The models were compared on the basis of their prediction capability, assessing the level of agreement between the measured data and predictions from the models.

To this aim, the monitoring network was partitioned into three sets of sites following these steps:

1. the 45 sites were stratified by type (urban/suburban, roadside and kerbside sites);
2. a random sample of nine sites was chosen, representative of the entire network (with respect to the number of sites of each type) as validation data for testing the models;
3. the other 36 sites were retained as training data to fit the models.

The steps (1)-(3) were repeated three times (so each site entered into the validation data once).

To evaluate the predictive performance of the models, the predicted PM_{10} concentrations were compared against the observed measurements on the validation set via the following indices:

- the empirical coverage of 90% credible intervals (90% CI)(that is essentially the proportion of time that the credible interval contains the observed value) coupled with their average length;
- the squared correlation coefficient (R^2);
- the root mean square error (RMSE) given by $\sqrt{\frac{1}{m}\sum_{l=1}^{m}\sum_{t=1}^{T}(y^*_{s_l,t} - y_{s_l,t})^2}$, where $y^*_{s_l,t}$ is the model predicted value of $y_{s_l,t}$ at time t at the validation site l, and m is the number of validation sites. Lower values of RMSE indicates more similarity among observed measurements and predicted values.

To obtain these indices, for each model we used the full posteriors from each Markov chain and we combined the predicted values from the three sets.

This same procedure was used to summarise the results for the parameters evaluation.

3.6.3 Computation

Computation was performed using the freely available Bayesian analysis software WinBUGS (Lunn et al. 2000), where BUGS stands for Bayesian Inference Using Gibbs Sampling. WinBUGS (that is the BUGS version working under Windows) uses the Gibbs sampling algorithm as Markov chain when possible (for example with conjugate distributions). In more complex situations, WinBUGS implements alternative algorithms, such as Metropolis sampling (Metropolis et al. 1953; Hastings 1970), slice sampling (Neal 2003) and various types of rejection method (e.g. for nonstandard but log-concave full conditionals, it uses the adaptive rejection sapling of Gilks (1992)).

Two parallel MCMC chains with different starting values were run for each model. We ran 60,000 iterations with 50,000 burn-in and thinned the Markov chains by a factor of 10, resulting in samples of size 2,000 to estimate the posterior distributions for the parameters of interest. Posterior correlation was reduced by a grand mean centring of the covariates (Gilks and Roberts 1996).

Convergence was assessed by checking the trace plots of the samples, the estimated kernel density plots, the autocorrelation functions, and a Monte Carlo errors <5% of the posterior standard deviation.

3.6.4 Predictions

Prediction in the Bayesian approach is based on the construction of a probability distribution of future values of the variable under study, conditional on the vector of past (observed) values, taking into account the posterior knowledge of the parameters. Thus, with the posterior computed, predictions are straightforward.

Let $\mathbf{y}^*(\mathbf{s}_0, t)$ be the PM concentrations at a set of unmonitored sites for the period in study that need to be predicted, and Θ be the collection of all parameters considered in the PM_{10} models, given the data, $\mathcal{D}_n$, the posterior predictive distribution of $\mathbf{y}^*(\mathbf{s}_0, t)$ is:

$$p(\mathbf{y}^*(\mathbf{s}_0, t)|\mathcal{D}_n) \propto \int p(\mathbf{y}^*(\mathbf{s}_0, t)|\mathcal{D}_n, \Theta) p(\Theta|\mathcal{D}_n) d\Theta \tag{3.30}$$

The spatial dependence was used to predict values on the spatial field (together with associated uncertainty) at locations where these were assumed not observed. Thus the spatial correlation matrix H was extended to include the new locations, that is: $\mathbf{y}(\mathbf{s}, t), \mathbf{y}^*(\mathbf{s}_0, t)|\Theta \sim \text{MVN}(\mathbf{0}, \omega^2 H(\phi))$.

3.6.5 Sensitivity analysis

Sensitivity analysis was performed in order to:

1. Assess the performance of our modelling approach in urban environments

that have a monitoring network less dense than in London. The EU Air quality directive (2008/50/EC) stipulates the minimum population dependent measurement requirements for EU cities. With 36 European cities with populations above one million and nine above two million (City Mayors Statistics 2012), we considered that testing the methodology on a sample of 10 measurement sites (matching the minimum number of monitoring sites for a city of 2.75 million population) would provide an assessment of applicability in a typical city. A city of 2.75 million would be smaller than the total area of Greater London. To this end, we considered the north-west boroughs in Greater London only and selected 10 sites as training set and three sites as validation set, representative of three site types, following the methodology described for the main analysis.

2. Corroborate the choice in modelling the long-range transport component of PM as latent variable. To this aim, model 1 was performed substituting the latent process with a simple predictor given by the average of the background measurements from the two monitoring rural sites, with fixed slope parameter. The typology of the two models was compared in term of predictive performance.

3. Compare the performance of the stochastically specification of the time-varying β_4 coefficient using a RW2 with a penalised spline specification. To this aims, knots were taken to be equally spaced over the range of the temperature variable and a parametric polynomial model was extended with the truncated polynomial basis functions (as specified in chapter 2), such that the form for smooth function for temperature assumed the form:

$$f(\text{temp}_t) = \delta_0 + \delta_1 \text{temp}_t + \sum_{h=1}^{H} \gamma_h (\text{temp}_t - \xi_h)_+^q \tag{3.31}$$

where $\delta_0, \delta_1, \gamma_1, \ldots, \gamma_H$ represents the vector of the regression coefficients, $(\text{temp}_t - \xi_h)_+$ are the basis functions, equal to $(\text{temp}_t - \xi_h)^q$ if $(\text{temp}_t - \xi_h)^q > 0$ and zero otherwise, and q is the degree of the spline, assumed to

be cubic. The vector $\delta = (\delta_0, \delta_1)$ represents the fixed effect coefficients, while the vector $\gamma = (\gamma_1, \ldots, \gamma_H)$ represents the random coefficients. Following Crainiceanu et al. (2005), the basis spline were constructed using radial basis functions (e.g., Ruppert et al. 2003). The idea behind the implementation of the penalised splines is to choose a generous dimension of the basis to achieve the desired flexibility, while the basis coefficients, γ, are penalised to avoid an overfitting of the data and ensure smoothness of the resulting functional estimates. In a Bayesian setting, as considered here, the coefficients γ were treated, as all the other parameters in the model, as random variables, and supplemented with a prior distribution. Thus, the penalisation was performed assuming as a prior for γ a Gaussian distribution with mean zero and variance σ^2_γ, to be estimated such that this variance component played a role of a smoothing parameter. In particular, a IG distribution with hyperparameters $(0.01, 0.01)$ was used.

4. Investigate whether results remained essentially unchanged in the presence of different hyperprior distributions. Commonly used inverse-Gamma priors were used for the variance parameters (measurement errors) $\sigma^2(\mathbf{s})$ and $\sigma^2_{lrt}(\mathbf{j})$: IG$(0.5, 0.0005)$ and IG$(0.1, 0.1)$. For the spatial effect variance parameter, ω^2, and the random walk variance parameter, σ^2_v, was tested the prior IG$(0.001, 0.001)$.

3.7 Results

3.7.1 Predictive performance

Table 3.1 shows the cross-validation summary statistics. The results are reported on the original scale.

Moving from model 1 to model 5, a progressive improvement in the prediction capability was observed, with exception of model 2. However, the validation indices improved heavily when the site-specific local component, described by

Table 3.1: Predictive performance by model (on original scale).

Models	Average width 90% CI	Coverage 90% CI	RMSE	R^2
Model 1	23.67	0.91	5.26	0.58
Model 2	45.55	0.88	11.11	0.04
Model 3	21.51	0.91	5.11	0.61
Model 4	22.20	0.89	5.04	0.61
Model 5	20.40	0.89	4.75	0.63

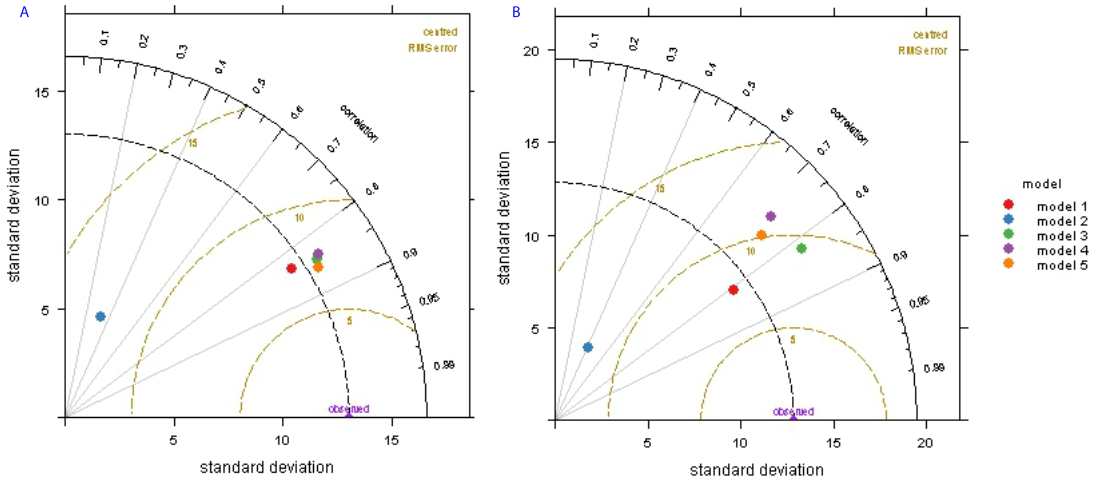
Abbreviations: CI, credible intervals; RMSE, root mean square error.

ADMS-Urban output, was included in addition to the regional background component (as an example, the RMSE decreased from 11.11 for model 2 to 5.11 for model 3). The incorporation of the selected covariates in models 4 and 5 produced an additional increase in the cross-validation performance.

Figure 3.6 shows the Taylor diagrams (Taylor 2001; Carslaw and Ropkins 2012) for the models, over (A) the whole study period and (B) a 2003's heat-wave event (days from 4th to 13th August 2003). This diagram represents a useful method for evaluating predictive performance, as it visualises simultaneously the centred RMSE (it is centred because the mean values of the observed and predicted data are subtracted first), the correlation coefficient (R) and the standard deviation of the observed and predicted values. In detail, the observed variability (i.e., the standard deviation) is plotted on the x-axis (specifically, the magnitude of the variability is measured as the radial distance from the origin of the plot), R is shown on the grey arc, while the RMSE is indicated by the concentric brown dashed lines emanating from the observed point.

The Taylor diagram performed on the entire study period (plot A) showed a quite similar performance of the models from 3 to 5, with model 5 be the best as presenting the highest correlation, the least RMSE and a reasonable similar variability compared to the observations, and also confirmed the poor performance of model 2. However, the Taylor diagram obtained on a 10 days heat-weave event (plot B) to assess how the models performed in capturing these events, pointed out differences, with models 2 and model 5 performing worst in comparison to models 1 and 3. This result could be explained by the fact that the heat-wave events of 2003 were dominated by the long-rang transport component.

Figure 3.6: Taylor diagrams showing the predictive performance of the five hierarchical models related to: (A) the entire period of study, and (B) a 2003's heat-wave event (from 4th to 13th August 2003).



Finally, Figure 3.7 shows the predictive performance of the five models according to site type, for several months of the year 2003. The sites plotted are chosen randomly within the sets used for the cross-validation. The superiority in performance of model 5 is clear for all the three site type sites, as well as the worse performance on model 2, missing pollution episodes and the August 2003 heat-wave.

3.7.2 Predictive performance of models implemented with varying intercepts

Table 3.2 presents the predictive ability of the models implemented using the classical approach given by space- and time-varying intercepts. Generally, the validation indices showed slightly worst values when compared with the cross-validation results from our modelling approach. However, for model III including the spatio-temporal random effects and the urban local component of PM, we found lower prediction errors in comparison to model 2 of our main analysis. This result confirmed that without temporal dependencies, the predictive capability of ADMS-Urban yielded poor performance.

Figure 3.7: Plots of observed PM_{10} concentrations (dots) and posterior means estimates (lines) by models for three different site type (A = Urban background; B = Roadside; C = Kerbside). Plots from March to September 2003.

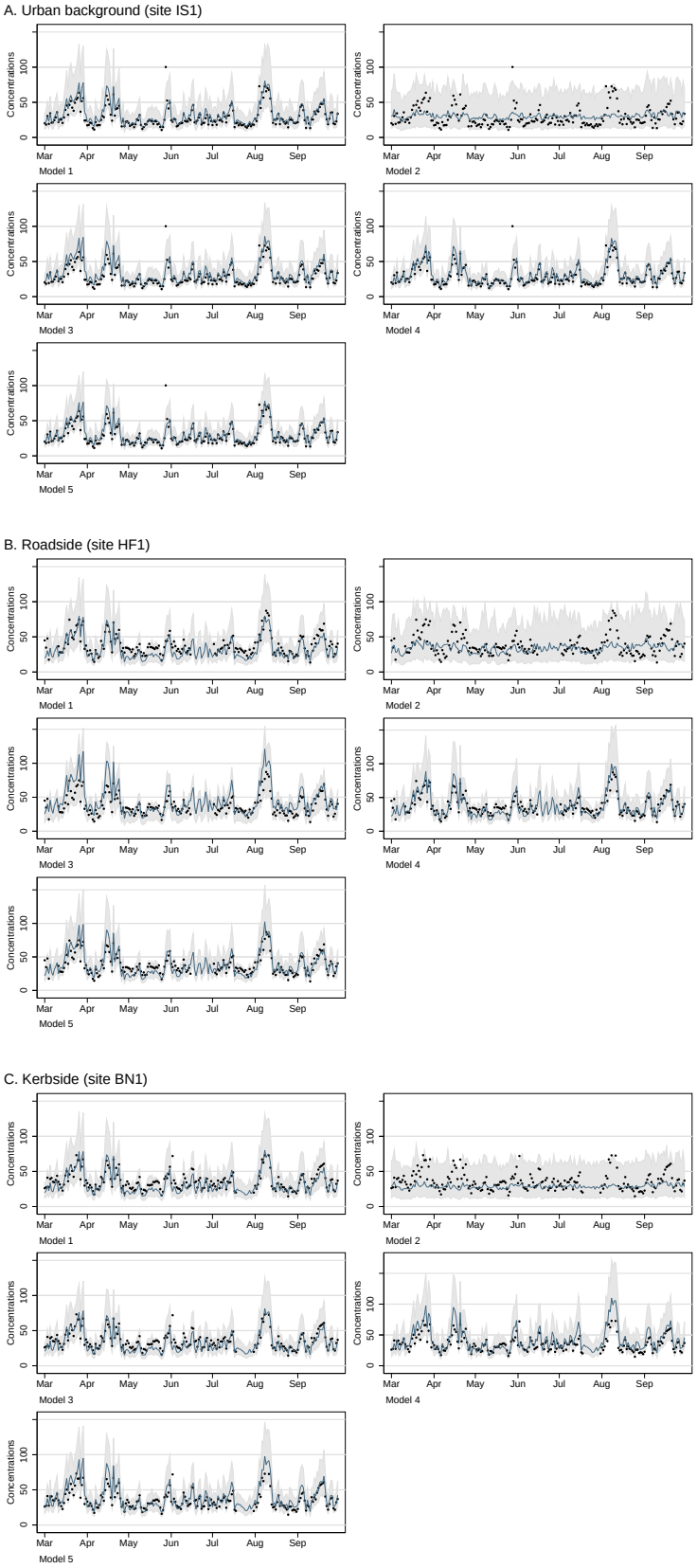


Table 3.2: Predictive performance of the models implemented using spatiotemporal varying intercepts (on original scale).

Models	Average width 90% CI	Coverage 90% CI	RMSE	R^2
Model I	28.58	0.89	7.37	0.64
Model II	29.90	0.88	7.58	0.64
Model III	28.43	0.92	6.84	0.65
Model IV	28.59	0.91	6.89	0.64
Model V	27.18	0.91	6.05	0.64

Abbreviations: CI, credible intervals; RMSE, root mean square error.

3.7.3 Parameter evaluation

The time-varying latent regional process described by $\mu_{lrt}(t)$ was found having a similar behaviour in models 1, 3 and 5. However, a visual inspection of the plot of the posterior mean of $\mu_{lrt}(t)$ pointed out a more evident daily variability in model 5. The range (on log-scale) of the posterior mean of the spatial coefficients, $\beta_1(\mathbf{s})$, associated with ADMS-Urban output, varied in model 2 from 0.005 to 0.333, in model 3 from 0.005 to 0.371, whilst in model 5 this ranged from -0.001 to 0.238. This suggested a weaker effect of the local PM_{10} component when the covariates were included in model 5.

Through the analysis of the decay parameter, ϕ, coherent results were found in models 2, 3 and 5, across all sets, for the spatial correlation among sites. Specifically, was detected a high correlation at minimum distance between sites, $\sim$0.97, that decayed progressively, being $\sim$0.50 at mean distance, and $\sim$0.24 at maximum distance.

Table 3.3 presents the posterior mean estimates and their 90% CI for the fixed effects and for the variance parameters.

The residual mean concentration, α, remained constant among the models. Instead, the variable site type described by β_2 played a strong effect, indicating that PM_{10} concentrations were greater for road and kerb sites than for suburban/urban sites, as expected. A negative relationship was estimated between PM_{10} and day of the week (described by β_3), as the concentrations were lower on the weekends than on weekdays.

The effect of the temperature on PM_{10} showed a considerable variability, espe-

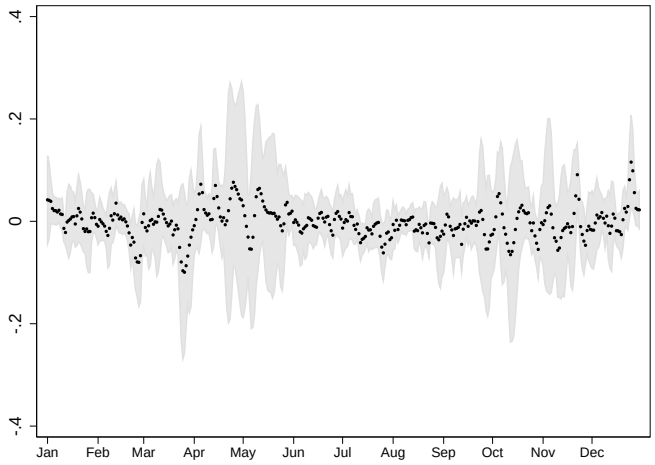
Table 3.3: Posterior mean and 90% credible intervals (CI) for the fixed effects and for the variance parameters by model (on log-scale).

Parameter	Model 1		Model 2		Model 3		Model 4		Model 5	
	Mean	90% CI	Mean	90% CI	Mean	90% CI	Mean	90% CI	Mean	90% CI
α (Intercept)	3.243	3.242, 3.244	3.315	3.309, 3.322	3.252	3.251, 3.253	3.325	3.302, 3,347	3.253	3.251, 3.254
$\beta_{2,2}$ (Road site)[a]	-	-	-	-	-	-	0.185	0.179, 0.192	0.143	0.142, 0.144
$\beta_{2,3}$ (Kerb site)[a]	-	-	-	-	-	-	0.282	0.269, 0.294	0.283	0.281, 0.284
$\beta_{3,2}$ (Saturday)[b]	-	-	-	-	-	-	-0.215	-0.276, -0.157	-0.032	-0.033, -0.030
$\beta_{3,3}$ (Sunday or Public Holiday)[b]	-	-	-	-	-	-	-0.201	-0.335, -0.074	-0.080	-0.082, -0.079
σ^2 (Range among sites of the posterior mean of variance)	0.061-0.202		0.163-0.168		0.038-0.0.074		0.048-0.052		0.033-0.050	
ω^2 (Spatial effect variance for the local PM component)	-	-	0.066	0.024, 0.152	0.041	0.040, 0.042	-	-	0.042	0.041, 0.043
σ_v^2 (Second order random walk variance for the temperature)	-	-	-	-	-	-	1.111	0.950, 1.300	0.006	0.005, 0.007

[a] Reference category: $\beta_{2.1}$ (Suburban/Urban site).
[b] Reference category: $\beta_{3.1}$ (Weekday).

cially for coefficients related to the spring days. Figure 3.8 shows the posterior mean of the time-varying coefficients, $\beta_{4,t}$ associated with temperature, for the year 2003. Time-varying coefficients indicate regional PM formations that was no fully accounted for by the measurements at Harwell and Detling and must therefore indicate secondary formation local to London. The TEOM instruments used were not sensitive to volatile PM components such as organics and ammonium nitrate (Green et al. 2009). These coefficients mostly likely represent ammonium sulphate formation. Greater variation in posterior mean during spring and autumn was consistent with increased emissions and therefore availability of agricultural ammonia at these times due to fertilizer use and manure spreading (Schaap et al. 2004).

Figure 3.8: Posterior mean estimates for the time-varying coefficients $\beta_{4,t}$ associated with temperature. Plot for year 2003.



Finally, with the exception of model 2, a progressive decrease in the measurement error variance across the models was noted. This reduction underlined the contribution given by the adjustment for covariates to explain part of the variability in the estimated PM_{10} concentrations.

3.7.4 Sensitivity analysis

Table 3.4 describes the results related to the predictive ability of the models on a restricted number of monitoring sites in north-west London. All the indices

were consistent with those reported in the main analyses (Table 3.1).

Table 3.4: Predictive performance by model obtained in the sensitivity analysis (on original scale).

Models	Average width 90% CI	Coverage 90% CI	RMSE	R^2
Model 1	31.52	0.93	6.91	0.52
Model 2	47.01	0.87	11.36	0.02
Model 3	29.62	0.92	6.65	0.57
Model 4	28.54	0.89	6.65	0.53
Model 5	23.29	0.88	5.38	0.61

Abbreviations: CI, credible intervals; RMSE, root mean square error.

The analysis performed to corroborate the choice in modelling the long-range transport component of PM as latent variable, compared to a simple inclusion of the rural measurements in the statistical model as additional predictors, confirmed the benefit of the choice (see Table 3.5).

Table 3.5: Predictive performance of two different statistical structure for model 1 (on original scale).

Model 1 formulations	Average width 90% CI	Coverage 90% CI	RMSE	R^2
$\mu(t,\mathbf{s}) = \alpha + \mu_{lrt}(t)$	23.67	0.91	5.26	0.58
$\mu(t,\mathbf{s}) = \alpha + \beta\,\bar{lrt}(t)$	30.65	0.92	7.13	0.33

Abbreviations: CI, credible intervals; RMSE, root mean square error. $\bar{lrt}$ represents the average of the background measurements from the two monitoring rural sites.

Moreover, a comparison of the predictive performance of the approach in using a RW2 priors for the time-varying coefficients β_4 associated with temperature, against a penalised cubic spline was performed. As expected, the results played similar results. In Table 3.6 are presented the results for model 5.

Table 3.6: Predictive performance for model 5 (on original scale) using a stochastic process RW2 and a penalised spline in modelling the time-varying coefficients for temperature

Method	Average width 90% CI	Coverage 90% CI	RMSE	R^2
RW2	20.40	0.89	4.75	0.63
Penalised B-splines	20.56	0.89	4.76	0.63

Abbreviations: CI, credible intervals; RMSE, root mean square error.

Finally, the evaluation of the sensitivity of findings to prior details, showed that the results were quite robust to these choices.

3.8 Discussion

This chapter presented a Bayesian spatio-temporal approach for modelling particulate pollution concentrations in urban area for short-term health risk studies.

The model combined air monitoring data with output from a local-scale air pollution model and explicitly solved the problem of incorporating regional pollution concentrations within the city scale assessment. The effect of covariates, included in the model to account for the residual spatio-temporal variation of particle concentrations, was also assessed. The evaluation of the predictive performance of these statistical structures was performed using a robust procedure of cross-validation that allowed the comparison of the daily predictions with the observed PM_{10} concentrations within three validation sets of sites, which represented different urban environment (i.e., site types).

In particular, the modelling approach was applied to enhance and to predict PM_{10} concentrations in Greater London, using a latent regional pollution process derived from rural sites to describe the long-range transport PM_{10} component and the output from ADMS-Urban to capture the local primary PM_{10} component.

ADMS-Urban is widely used for estimating urban scale air pollution for regulatory purposes and in epidemiologic air pollution studies (e.g., Laurent et al. 2008; Blangiardo et al. 2011). From this analysis, it is clear that the exclusive use of ADMS-Urban to predict the PM_{10} concentrations produces poor results. So far, although the inclusion of ADMS-Urban, in addition to a regional latent process, increases the predictive performance of the models, the study suggests that the use of this deterministic output to measure the population exposure to particle pollution in short-term epidemiologic studies, should be enhanced with the combination of other information sources characterising the study area, such as site type or time-varying emission factors linked to day of the week, as evidenced by the strength of the covariates in our models.

In this implementation, the indicator variable adopted for site types was actually quite crude. The use of a more localised index of sites better reflecting

land use and building geometry (canyon orientation for example) by utilising GIS techniques may further improve the model performance. Moreover, the long-range transport component of PM was described by a latent variable based on measurements of two rural monitoring sites. This aspect can be further improved by using more sites, selecting them according wind directions. This would ensure that the latent variable is able to better describe the PM concentrations up winds to London (Bressi et al. 2013).

The final goal of this study was to perform air particle pollution exposure models to use in short-term health effects studies in London. Therefore the work was developed with the dense monitoring network available in due to the city size and the legal structures for local air quality management. To assess the applicability of this approach in urban environment with smaller number of monitoring sites, a sensitivity analysis was performed, restricting the study area to a part of London matching the minimum requirements in EU directives. The results suggested that the approach will also perform well in smaller urban environments with more sparse monitoring networks, which are typical of many European cities.

Methodologically, the models presented here deviate from the standard space-time statistical modelling approach which typically presents varying intercepts (e.g., Shaddick and Wakefield 2002; Sahu et al. 2006; Cocchi et al. 2007; Sahu et al. 2009; Berrocal et al. 2010b). However, as specified in Gelman and Hill (2007) there are situations in which a constant intercept and varying slopes model can be reasonable. As the models included variables characterised by spatial and temporal variation, thus only time- and space-varying regression coefficients were assumed. To assess the plausibility of this approach in comparison to a classical modelling scenario, five models characterised by independent spatio-temporal random effects were performed. Assessing the predictive capability of these structures, was clear that the adopted methodology, applied in an urban environment, performed better than the classical approach. This evidence suggests that, in context where local and urban primary emissions together with regional back-

ground data are not available, the inclusion in the models of independent error distributions is able to capture spatial and temporal dependencies. However, in context of analysis, where the researchers can perform extra modelling efforts, the approach here proposed performs better than a classical approach.

Finally, the hierarchical methodology proposed provided a flexible way to model daily particle pollution. This approach could also be applied to other environmental space-time processes (e.g., to model time-series of different ambient primary or secondary pollutants) and used to predict non-daily data (e.g., hourly).

4 | Health effects of exposure to temporal airborne particle profiles

The core of this chapter is represented by a Bayesian semiparametric modelling approach, defined by a DP mixture model, for estimating the association between PM characteristics and short-term health effects, with the objective of providing new insights into the identification of the differential harmful effect of PM based on its components and sources. In chapter 2 was pointed out that multipollutant epidemiologic time series studies need to deal with the problem of highly correlated nature of the exposure metrics. The approach here presented is specifically conceived to examine the joint effect of different metrics on an health outcome, overcoming several limitations of traditional regression methods.

Section 4.1 presents the background for the development of the model described in this chapter. Then, section 4.2 briefly recalls the basic concepts of DP mixture modelling, with emphasis on stick breaking process construction of the DP. Section 4.3 illustrates the daily data used for the study, consisting of a range of particle metrics and respiratory mortality for London 2002-2005, as well as the particle metrics for the year 2012 used as new exposure scenario to predict mortality, and the confounding factors included in the analysis. Sections 4.4 and 4.5 present the model and the results, along with the predictions and the sensitivity analyses. For comparison, an application using more conventional statistical tools is also presented. This relies on a two-step procedure involving the popular K-means algorithm for the clustering problem and a regression model to study the effect of grouped mixtures on respiratory mortality. Section 4.6 presents the methods and the results for the comparative methodology. Finally, Section 4.7

presents strength and limitations of Bayesian profile regression for air pollution health studies in a time series design.

The work presented in this chapter, obtained using the Bayesian DP mixture model, is based on a recently published peer-reviewed article in *Environment International* by Pirani et al. (2015). The paper was coauthored by Nicky Best, Marta Blangiardo, Silvia Liverani, Richard W. Atkinson and Gary W. Fuller, who provided data sets, supervised the analyses and contributed to the interpretation of the results.

4.1 Background

Most of the studies on air pollution time series have focused on the health risk associated with the total mass of particles, without considering the heterogeneity in their chemical and physical composition. In the last decade, however, there has been a growing interest, in environmental research communities, in assessing the health effects of simultaneous exposure to different particle pollutants. Moreover, policy-makers would benefit from information on which components or sources are most harmful.

In a recent review of techniques for characterising air pollution exposure metrics, Oakes et al. (2014) highlighted that clustering of air pollutant profiles has been shown to be an effective approach. Temporal clustering analyses have been successfully applied in air pollution exposure assessment, involving mainly standard heuristic methods such as agglomerative hierarchical clustering (Gu et al. 2012) and K-means partitioning clustering (Austin et al. 2012). Recently, K-means clustering solutions of air pollutants have also been used as covariates within health model effect estimation (Matyasovszky et al. 2011; Zanobetti et al. 2014). Other literature contributions in air pollution have proposed mixture models as alternative technique to heuristic clustering methods (Gómez-Losada et al. 2014).

Mixture models in both their structures of finite or infinite number of components represent an attractive methodology for clustering, as they bring back the problem to a probabilistic domain. The modelling approach presented in this chapter is an infinite mixture model within the Bayesian framework and it aims to learn about the simultaneous impact of multiple particle metrics on health outcomes, using the DP mixture model defined by a stick-breaking construction. The statistical background is included in the following sections. In particular, this approach is based on a class of methods in Bayesian nonparametrics in which the DP mixtures is used to model the joint distribution of the response and covariates. Müller et al. (1996) firstly introduced a joint distribution for response and covariates. Further studies include the works of Taddy and Kottas (2009); Kang and Ghosal (2009); Shahbaba and Neal (2009); Müller and Quintana (2010) and Wade et al. (2014). Hannah et al. (2011) proposed a DP mixtures of GLMs.

The model presented here builds on the recent work of Molitor et al. (2010, 2011), and it represents an alternative inferential approach to regression models when the covariates under study are correlated, as it happens for particle pollution. This technique, known as *profile regression*, performs a DP Bayesian clustering of the covariates by identifying exposure profiles and, simultaneously, links these to a response variable in non-parametric form (even though the model continues to be parametric within clusters). Profile regression has also been applied in epidemiology and in genomics (Papathomas et al. 2011, 2012; Hastie et al. 2013).

Here this model is extended to analyse time series data, accounting for their typical features like trend, seasonality and temporal components through smooth functions. The resulting probabilistic solution groups time points with similar multipollutant and response profiles.

To demonstrate the approach, a data set of daily particle metrics from London 2002-2005 and daily number of deaths from respiratory diseases was used (Atkinson et al. 2010). Additionally, to assess changes in respiratory mortality from

the recent efforts in reducing air pollution in London, a mean response profile for mortality in the year 2012 was predicted, using as exposure scenario particles measured at the same monitoring site. Thus, a comparison of the predictive distribution of mortality in 2012 against the one computed in 2005 was performed.

To illustrate the strengths of the proposed model as well as the differences between this and other clustering techniques, an application using the common K-means clustering is also presented. K-means is a classic example of a non-probabilistic vector clustering method that uses iterative relocation, tempting to minimize the within-cluster variance (see chapter 2), section 2.3.8. The analysis of the association between clustering solution and mortality is performed accordingly to Zanobetti et al. (2014).

4.2 Preliminaries on Dirichlet process and infinite mixture models

The DP (Ferguson 1973, 1974; Blackwell and MacQueen 1973) is a stochastic process which defines a distribution over distributions. It can be view as an extension of the the *Dirichlet distribution* to continuous spaces.

4.2.1 Dirichlet distribution

The Dirichlet distribution (e.g., Murphy 2012) is a multivariate generalisation of the Beta distribution. Let $\mathbf{x}$ be a K-dimensional vector. The Dirichlet distribution has support over the probability simplex, defined by:

$$S_K = \left\{ \mathbf{x} : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1 \right\} \tag{4.1}$$

The Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ can be written as follows:

$$p(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \tag{4.2}$$

where $\Gamma(\cdot)$ is the Gamma function. The first term on the right side of (4.2) is the multinomial Beta function that serves as the normalising constant. The K-dimensional vector $\boldsymbol{\alpha}$ has components $\alpha_k > 0$, and determines the variance in the values of $\mathbf{x}$. Their sum $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ can be interpreted as a precision (or concentration or scale) parameter. If $K = 2$ the Dirichlet distribution reduces to the Beta distribution.

The mean and variance of the Dirichlet distribution are respectively:

$$E(x_k) = \frac{\alpha_k}{\alpha_0} \tag{4.3}$$

$$\mathrm{Var}(x_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \tag{4.4}$$

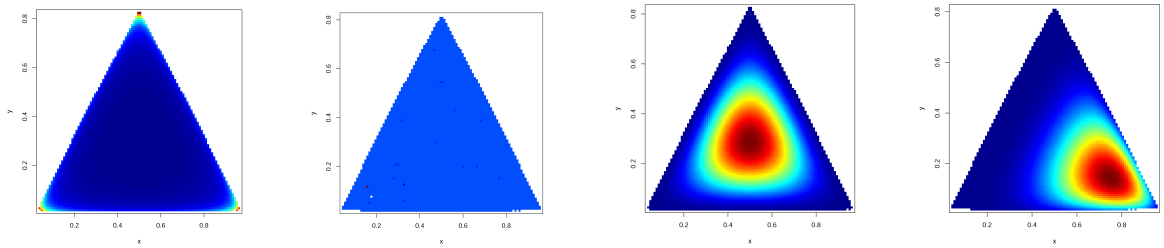
The result of sampling from a Dirichlet distribution is itself a distribution on some discrete probability space. In Bayesian statistics, the Dirichlet is a conjugate prior distribution for the parameters of the Categorical distribution (that generalises the Bernoulli distribution to more that two states) and for the Multinomial distribution.

Figure 4.1 shows the probability density function of Dirichlet distribution for various $\boldsymbol{\alpha}$ parameters, with the simplex projected into two dimensions. Red shades correspond to high probability and blue shades correspond to low probabilities. When all values of $\boldsymbol{\alpha}$ are set to 1, this corresponds to a uniform distribution over the simplex, that is all the points in the simplex have the same probability (plot (b)). If $\boldsymbol{\alpha} < 1$ there are sharp peaks of density almost at the vertices of the simplex and if $\boldsymbol{\alpha} > 1$ the density becomes concentrated toward the center of the simplex.

4.2.2 Dirichlet process

The DP is like an infinite-dimensional Dirichlet distribution, where each draw returns a distribution F over a countably infinite set of outcomes. It is used in Bayesian nonparametrics to represent the uncertainty about the parametric form of a distribution.

Figure 4.1: Density plots (blue = low probability, red = high probability) for the Dirichlet distribution over the probability simplex in $\mathbb{R}^2$ for various values of the parameter $\boldsymbol{\alpha}$; (a): (0.1, 0.1, 0.1), (b): (1, 1, 1), (c): (3, 3, 3), (d): (5, 2, 2).



(a) (b) (c) (d)

The DP was formally defined by Ferguson (1973), who proposed its use as a prior over the set of all probability distributions on a given sample space. Let $(\Theta, \mathbb{A})$ be a measurable space. Suppose F_0 is a probability distribution (measure) with support in space Ω and α a positive real number. Thus, F is distributed according to the DP with base distribution F_0 and concentration parameter α, if for any finite measurable partition $(A_1, \ldots, A_K)$ of Θ, with $A \in \mathbb{A}$, a random vector $(F(A_1), \ldots, F(A_K))$ is distributed as a finite-dimensional Dirichlet distribution:

$$(F(A_1), \ldots, F(A_K)) \sim \text{Dir}(\alpha F_0(A_1), \ldots, \alpha F_0(A_K)) \tag{4.5}$$

Ferguson (1973) proved the existence of this process and showed that F is discrete asymptotically. In notation, $F \sim \text{DP}(\alpha, F_0)$ is used to denote that the random probability measure F follows a DP.

The DP holds several important properties that are as follows:

- The base measure F_0 defines the mean of the process, and for any measurable set $A \subset \Theta$, it is $E[F(A)] = F_0(A)$; while the α parameter can be understood as an inverse variance: $\text{Var}[F(A)] = F_0(A)(1 - F_0(A))/(\alpha + 1)$ (Teh 2010). So the larger α is, the smaller the variance (broadly speaking, α controls the number of components of the mixture).

- Given a set of independent observations $\theta_1, \ldots, \theta_N$ from F, the posterior

distribution is also a DP (Ferguson 1973):

$$F|\theta_1, \ldots, \theta_N \sim DP\left(\alpha + N, \frac{\alpha}{\alpha + N} F_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}\right) \tag{4.6}$$

In particular, the concentration parameter becomes $\alpha + N$ after observing N samples, and the contribution of the prior base distribution F_0 is scaled by α.

- The base measure F_0 is continuous, so the probability that any two samples are equal is precisely zero. However, Blackwell (1973) showed that F is a discrete distribution, made up of a countably infinite number of point masses. Therefore, there is always a non-zero probability of two samples colliding.

There are many ways to construct the DP. Here the attention is given to the *stick-breaking process*, as it is the construction used in modelling approach presented in this chapter.

4.2.3 Stick-breaking process

The construction of the DP as the stick-breaking process is due to Sethuraman (1994). He characterised the DP realisations as a countable mixture of point masses. According this definition, a random probability distribution, F, generated from a DP is (almost surely) of the form:

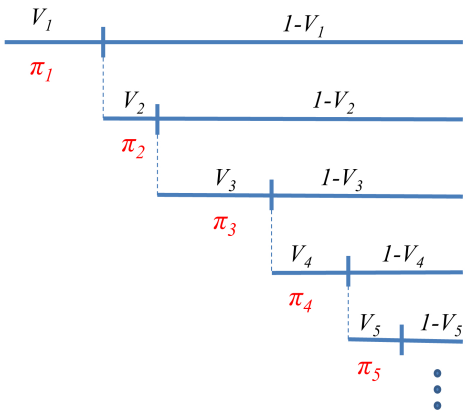
$$F \stackrel{d}{=} \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \tag{4.7}$$

where δ_θ denotes a Dirac measure (point mass) at θ. The locations of the point masses, $\{\theta_1, \theta_2, \ldots\}$, are i.i.d sample from F_0.

The probability weights, π_k, arise from a stick-breaking process. The name of this construction derives by an analogy given by breaking pieces off from a stick of unit length, where the breakpoints are randomly sampled from the Beta distribution. The mixture probabilities break the stick into a potentially infinite number of

pieces, such that $\sum_{k=1}^{\infty} \pi_k = 1$. Let $V_1, V_2, \ldots$ be random variables independent of θ's and i.i.d. among themselves with common distribution Beta$(1, \alpha)$. Thus, the first mixture probability is equal to V_1, that is $\pi_1 = V_1$, where $V_1 \sim$ Beta$(1, \alpha)$, and for $k \geq 2$ the kth mixture probabilities are given by $V_k \prod_{l=1}^{k-1}(1 - V_l)$. Figure 4.2 illustrates graphically the stick-breaking process.

Figure 4.2: Graphical representation of the stick-breaking construction of the Dirichlet process (modified from Ghahramani (2005)).



The stick-breaking distribution over π is sometime written $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$, where GEM stands for Griffiths, Egen and McCloskey, to indicate a set of mixture weights sampled from this process (Pitman 2006; Teh 2010).

4.2.4 Dirichlet process for mixture models

One of the most important applications of the DP is as a prior on the parameters of a mixture model. As discussed in chapter 2 (section 2.3.8), mixture distributions are a tool for modelling processes whose output is thought to be generated by different underlying mechanisms, or to come from different populations (Neal 1992). In a finite mixture model (Richardson and Green 1997; McLachlan and Peel 2000), the data are modelled by a finite but unknown number (K) of probability distributions; Bayesian nonparametric mixtures, instead, use mixing distributions consisting of a countably infinite number of components (Orbanz and Teh 2010).

Let a set of observations, $x_1, \ldots, x_N$ be modelled by a set of latent parameters

$\theta_1, \ldots, \theta_N$, $i = 1, \ldots, N$. Each θ_i is drawn independently and identically from F, while the data points x_i are i.i.d. with distribution function $f(\theta_i)$:

$$\begin{aligned} x_i|\theta_i &\sim f(\theta_i) \\ \theta_i|F &\sim F \\ F &\sim \mathrm{DP}(F_0, \alpha). \end{aligned} \tag{4.8}$$

where F is the unknown random distribution over parameters generated from a DP. Because F is discrete, multiple θ_i's can take on the same value simultaneously and the model in (4.8) can be seen as a mixture model, where x_i's with the same value of parameter ($\theta_i = \theta_j$) belong to the same cluster (Teh 2010).

4.3 Description of the data

Atkinson et al. (2010) described results from an epidemiologic air pollution health effect time series study examining the effect of different metrics of particulate collected in London, on cardiorespiratory hospital admission and mortality using univariate log-linear Poisson models. In the work proposed here, a subset of exposure data for the period January 2002 to December 2005 (years 2000-2001 were excluded due to pour data availability; for anions the proportion of missing data was about 96%), and respiratory-related mortality as the outcome was selected. To predict respiratory mortality, within the Bayesian profile regression model, given the multipollutant scenario that London experienced in 2012, the same set of particle metrics that were recorded in 2002-2005 was measured in 2012.

4.3.1 Mortality data

Daily count of deaths from respiratory diseases of London residents (2002-2005) were obtained from the Office for National Statistics and coded using the International Classification of Diseases, 10th Revision (ICD−10: Chapter J).

4.3.2 PM measurements 2002-2005

Daily average concentrations of particle metrics included: PNC, inorganic anions such as chloride, nitrate and sulphate[1], black smoke (BS) and gravimetric measurements of PM, such as PM_{10}, $PM_{2.5}$ and $PM_{10-2.5}$ (coarse fraction was obtained for subtraction).

With the exception of BS, the daily concentrations were obtained from a single background monitoring station in central London (North Kensington). BS was an average across several urban and suburban stations. PNC was measured using a TSI 3022A condensation particle counter, where particles are enlarged by condensation of saturated butanol vapour which are then counted using a laser and optical detector. The PM_{10} 24-hour filter samples were collected at 16.7 l per minute on quartz fibre filters using Partisol 2025 (Thermo) instruments and these filters were analysed by ion chromatography. Finally, daily average gravimetric PM_{10} and $PM_{2.5}$ were sampled using a Partisol sampler and measured using methods in EN12341 and EN14907.

The data set also included PM apportioned into primary and non-primary sources (Fuller et al. 2002; Fuller and Green 2006), giving modelled primary PM_{10} (PPM_{10}), and non-primary PM subdivided by size fraction: non-primary PM_{10} ($NPPM_{10}$), non-primary $PM_{2.5}$ ($NPPM_{2.5}$), and non-primary PM coarse fraction (NPcoarse). The source apportionment model assumed that primary PM_{10} was associated with NO_x sources and the non-primary component was the fraction of PM not associated with NO_x. NO_x is generally considered a robust marker for traffic pollution (Krzyzanowski et al. 2005).

4.3.3 PM measurements 2012

The ambient PM concentrations for the year 2012 were exclusively used within the Bayesian profile regression model to predict the mortality counts for respiratory-related diseases.

[1]For ease of clarity, in this chapter anions are defined by their name and not by their chemical formula.

The PM measurements (except BS) were collected at the same background monitoring site in central London. Between 2005 and 2012, gravimetric filter substrates were changed from quartz fibre to PTFE coated glass fibre (Emfab, Pall). Because BS is no longer measured in London, the daily mean of BS were obtained from equivalent measured black carbon by aethalometer (Magee Scientific) at two background monitoring sites in London, and an adjustment factor of 0.27 was applied following Heal and Quincey (2012).

4.3.4 Confounding factors

Ecological time series studies are subject to complex forms of confounding (e.g., Peng et al. 2006; Bhaskaran et al. 2013). In chapter 2 has been pointed out that, typically, time series studies of mortality and morbidity control for long-term trends, seasonality, and time-varying factors, including meteorological variables, which can potentially confound the association between an adverse health effect and polluted air. In this work, calendar time and temperature were considered as confounding variables and assumed to potentially influence the response variable via smooth functions.

Specifically, for all of the smooth functions were used natural cubic spline bases. As seen in chapter 2 (section 2.1.1) splines are flexible models that take the form of piecewise polynomials joined at knots and continuity constraints are generally imposed at the knots so that the function is smooth. Natural cubic splines have continuous first and second derivatives at the knots, but the second derivatives at the two end-points are taken as zero. A meaningful measure of the amount of smoothing is given by the *effective degrees of freedom* (Buja et al. 1989), commonly shortened to *degrees of freedom* (*df*). For natural cubic splines the *df* are equals to the number of knots plus 1 (plus intercept).

In this study, the choice of the *df* was based on the examination of the partial autocorrelation function of residuals and by minimization of the Akaike's Information Criterion (AIC) (Akaike 1973) and the Bayesian Information Criterion (BIC) (Schwarz 1978), fitting a log-linear Poisson regression model.

For the smooth function of time 32 *df* (8 *df* per year) were specified and for the smooth function of temperature 3 *df*. In a previous study performed to investigate the potential for bias in estimating the short-term effects of air pollution, Shaddick et al. (2013) used London's 2002-2005 respiratory mortality and PM_{10} concentrations and showed that a similar adjustment provided an adequate balance between ensuring control for temporal trends and seasonal cycles as well as temperature, while leaving sufficient information for estimating the exposure effects.

For this study, hourly temperatures were downloaded from the London Air Quality Network using the R library *openair* (version 0.9-2) (Carslaw and Ropkins 2012) and averaged on daily temporal scale. In particular, temperature data have been compiled from three meteorological sites across London with the aim to represent a 'typical' conditions in the city. During the years 2002-2005, daily average temperature ranged from -0.88 °C to 28.87 °C.

For the Bayesian profile regression model, the B-spline basis matrix for the natural cubic splines of calendar time and temperature were generated outside the model, using the function *ns* of the R library *spline*, and entered as data.

4.3.5 Data processing

The exposure data were normalised to be on a comparable scale adopting the modified z-score recently proposed by Austin et al. (2012). Let $x_{t,j}^{orig}$ be the original measurement on day t of particle metric j, for $t = 1, \ldots, T$ and $j = 1, \ldots, P$. The original measurements were transformed as:

$$x_{t,j} = (x_{t,j}^{orig} - \text{Median}(x_j^{orig}))/(\text{Median}(|x_{t,j}^{orig} - \text{Median}(x_j^{orig})|)) \qquad (4.9)$$

In a previous analysis, Atkinson et al. (2010) observed associations for respiratory mortality with 1−day lag secondary PM masses. The estimated regression coefficients were obtained fitting separate univariate log-linear Poisson models.

To study the value added by this new approach, the previous study of Atkinson et al. (2010) was considered as benchmark and thus the 1 day lag was chosen as the exposure window for particles.

4.4 Bayesian profile regression

4.4.1 Model specification

As previously discussed, the proposed model is based on the DP, which relies on mixtures to represent distributions in the data.

Denote by $t = 1, \ldots, T$ a series of temporal points. Let the data consist of realizations of a response data vector $\mathbf{y} = (y_1, \ldots, y_T)$, a set of (normalised) covariates (i.e., predictors) $x_{t,j}$, $j = 1, \ldots, P$, and a collection of confounding factors $u_{t,l}$, $l = 1, \ldots, L$, occurring according to some underlying random joint distributions. In this study, y_t denotes the count number of deaths for respiratory diseases on day t, $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,P})'$ represents a daily covariate profile of air particles, and $\mathbf{u}_t = (u_{t,1}, \ldots, u_{t,L})'$ represents a daily observation for the confounding factors, i.e., calendar time and temperature (that is, $L = 2$).

A joint probability model for the data is assumed, which takes the following form:

$$p(y_t, \mathbf{x}_t|\Theta, \mathbf{u}_t) = \sum_{k=1}^{\infty} \pi_k p(y_t|\Theta_k, \Theta_0, \mathbf{u}_t) p(\mathbf{x}_t|\Theta_k, \Theta_0) \tag{4.10}$$

where π_k are the mixture probabilities satisfying $\sum_{k=1}^{\infty} \pi_k = 1$ almost surely and indicating the probability of belonging to the kth component. Θ denotes the collection of model parameters, that includes component specific parameters, Θ_k, and global parameters, Θ_0, that is, $\Theta = (\Theta_k, \Theta_0)$.

As seen in chapter 2 (section 2.3.8), the inference for such mixture models can be simplified by introducing latent variables that indicate the group memberships of objects (i.e., the cluster to which day t belongs to). Let $\mathbf{z} = (z_1, \ldots, z_T)$ be the latent group labels, such that $p(z_t = k) = \pi_k$. Thus, z_t is chosen using a Multinomial distribution parameterised by the mixing probabilities, $z_t|\boldsymbol{\pi} \sim \text{Mult}(\boldsymbol{\pi})$.

The mixture weights, π_k, are generated using a stick-breaking procedure (Sethuraman 1994), based on i.i.d. Beta distributions, as described in section 4.2.3. A Gamma distribution was used to specify prior uncertainty for the precision parameter, α, of the DP (following Escobar and West (1995)), namely $\alpha \sim \text{Gamma}(a, b)$, where $a = 2$ and $b = 1$ are the shape and the inverse-scale (rate) parameter respectively.

A multivariate Normal distribution for the P covariates was assumed:

$$p(\mathbf{x}_t|\Theta_k, \Theta_0) = (2\pi)^{-\frac{P}{2}}|\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_t - \mathbf{m}_k)'\Sigma_k^{-1}(\mathbf{x}_t - \mathbf{m}_k)\right\} \tag{4.11}$$

where $\mathbf{m}_k = (m_{k,1}, \ldots, m_{k,P})$ is the mean vector for component k (i.e., location parameters), and Σ_k is the P x P symmetric and positive-definite variance-covariance matrix.

The hyperpriors for $\mathbf{m}_k$ and Σ_k were specified similarly to Molitor et al. (2011), adopting an empirical Bayesian approach. A Normal distribution was assumed for the location parameters, that is, $\mathbf{m}_k \sim \text{N}(\mathbf{m}_0, \Sigma_0)$ (with $\mathbf{m}_0$ equal to the empirical mean of each covariate, and Σ_0 having a diagonal structure with elements equal to the square of empirical range of each covariate). A Wishart distribution was specified for the precision matrix Σ_k^{-1} (i.e., inverse variance-covariance matrix), that is, $\Sigma_k^{-1} \sim W(\Phi, \nu)$, where Φ is a symmetric and positive-definite matrix parameter (set equal to the inverse of the empirical covariance matrix multiplied by $1/P$) and ν is the degrees of freedom parameter (set equal to P).

The response was modelled as a Poisson:

$$p(y_t|\Theta_k, \Theta_0, \mathbf{u}_t) = \frac{\lambda_t^{y_t}}{y_t!}\exp(-\lambda_t) \tag{4.12}$$

where

$$\lambda_t = E_t \exp(\mu_t) \tag{4.13}$$

and

$$\mu_t = \mu_k + f_1(t) + f_2(\text{temp}_t) + \epsilon_t \tag{4.14}$$

assuming ϵ_t to be normal distributed with zero mean and variance σ^2_ϵ. Here μ_t is the mean response for day t and E_t is the expected offset given by the average number of deaths for respiratory diseases in the full period in study.

The parameter of interest is μ_k, which represents the log relative risk for the outcome of interest associated with the kth cluster. Each cluster includes days with similar multipollutant profile. The confounding factors of time and temperature, previously collectively represented as $\mathbf{u}_t$, are assumed to vary smoothly, having a natural spline representations as specified in (4.14). In detail, the functions $f_1(\cdot)$ and $f_2(\cdot)$ denote smooth functions of time and temperature respectively:

$$f_1(t) = \sum_{i=1}^{n_1} \gamma_{1,i} B_i(t)$$
$$f_2(\text{temp}_t) = \sum_{i=1}^{n_2} \gamma_{2,i} H_i(\text{temp}_t)$$

Here, n_1 and n_2 are the *df* for f_1 and f_2 respectively and B_i and H_i are the basis functions. The relative coefficients $\gamma_{1,1}, \ldots, \gamma_{1,n_1}$, $\gamma_{2,1}, \ldots, \gamma_{2,n_2}$ are assumed to follow a weakly informative Student-t prior distribution, with location, scale and degree of freedom set to 0, 2.5 and 7 respectively (Gelman et al. 2008). The smooth functions were constrained to only have a global effect on the response and not a cluster-specific effect.

4.4.2 Computation

Inference for this model relies on MCMC computational methods. A slice dependent sampler algorithm for posterior computation was used, as implemented in the R package *PReMiuM* (version 3.0.24) (Liverani et al. 2015). Slice sampling methods go back to Neal (2003), and have been successively described for DP mixture models by Walker (2007) and Kalli et al. (2011). The basic idea is to introduce an auxiliary latent slice variable that allows a finite number of clusters

to be sampled within each iteration of the sampler. The algorithm implemented in *PReMiuM* combines a Gibbs sampler with Metropolis-within-Gibbs steps. It also implements label switching moves as suggested by Papaspiliopoulos and Roberts (2008). Details regarding this sampler are given by Liverani et al. (2015).

The algorithm was run for 70,000 iterations with the first 20,000 discarded as burn-in. Using 1 in 10 thinning, this gave us a total of 5,000 draws from the posterior distribution of parameters and predictions.

Convergence was checked through the inspection of trace plots of the samples, the estimated kernel density plots and the autocorrelation plots of the main global parameters of the model using the R package *coda* (version 0.16-1).

4.4.3 Post-processing

The estimation task in the Bayesian model-based clustering is complicated by the label switching. It means that during the MCMC run the labels associated with the clusters change. To summarise the features of the rich output from the MCMC sampler, a post-processing of the posteriors was performed, as suggested by Molitor et al. (2010, 2011), that relied on a *representative* partition (i.e., that is most supported by the data) obtained by using a similarity matrix based upon the output of the MCMC.

In particular, independently of any labelling, at each iteration of the sampler, a pairwise cluster membership was recorded and a T x T score matrix was constructed, with (i, j)th element set equal to 1 if day i and day j belong to the same cluster and 0 otherwise. The end of this process leads to a probability matrix, $\mathbf{S}$, formed by averaging the score matrices obtained at each iteration, thus element $S_{i,j}$ denotes the probability that day i and j are assigned to the same cluster. A clustering procedure PAM (Kaufman and Rousseeuw 1990) was used on the dissimilarity matrix $1-\mathbf{S}$ to obtain representative partitions.

Once the representative clustering was defined, a model averaging approach was adopted to evaluate the uncertainty related to the characteristics of the clusters that involved running through the MCMC run, obtaining an average value for

the model parameters (effects and cluster related parameters) across all days in a certain cluster.

4.4.4 Cross-validation and predictions

Classical regression provides concentration response functions that can be used in health impact assessment or to assess the costs and benefits of policies to decrease pollution exposures. By using profile regression, the types of daily pollutant mixtures that were associated with adverse health effects were identified and quantified. It also allowed to analyse what would happen to this health outcome if the exposure variables were changed. This was accomplished by a predictive approach (Müller et al. 1996).

The main idea here was to obtain a posterior predictive distribution of the response, given a new exposure scenario. In our application the simulated predictions represented an average effect of the changed air particle mixtures in London.

Two predictive scenarios were compared based on: (i) concentrations of particles measured in 2005, and (ii) concentration of the same particles measured in 2012, to analyse any changes in respiratory mortality arising from the combined effects of local, city, national and EU policies to manage air pollution in interval of seven years period.

The posterior predictions were carried out using the method proposed by Liverani et al. (2015) using simple allocations where, at each sweep r of the MCMC sampler, is assigned $\hat{\mu}_s^r = \mu_k^r$. In particular, an additional latent indicator variable $\hat{z}_s^r$ was defined, corresponding to each predictive scenario. Let $\mathbf{x}_t^* = (x_{t,1}^*, \ldots, x_{t,P}^*)$ be the new profile of exposure, the posterior probabilities were computed as:

$$p(\hat{z}_s^r = k | \mathbf{x}_t^*, \Theta^r, y, \mathbf{x}_t) \tag{4.15}$$

Given these probabilities, a predicted averaged cluster-specific estimate of the

response is performed, for each new profile of particles at each sweep:

$$\hat{\mu}_s^r = \sum_{k=1}^{\infty} p(\hat{z}_s^r = k|\mathbf{x}_t^*, \Theta^r, y_t, \mathbf{x}_t)\mu_k^r. \tag{4.16}$$

Before computing predictions of mortality for the different exposure scenarios, a predictive cross-validation technique was used as model checking. The four-years time series was partitioned, using the data collected in 2002-2004 as training sample and the data in 2005 as validation sample. Respiratory deaths were predicted for the 2005 and subsequently the validation predictions were compared with the actual observations. Specifically, the observed mortality was compared with the validation predictions in the year 2005 using the adjusted R^2 and the root mean squared error (RMSE) given by $\sqrt{\frac{1}{T_v}\sum_{t=1}^{T_v}(y_t^* - y_t)^2}$, where T_v is the number of observations for the validation set (i.e., 365 days), y_t^* and y_t are respectively the predicted and observed mortality.

Then the full four-years time series data were used for the computation of the posterior predictive distribution of the count of respiratory-related deaths in 2012, and this was compared with the one computed for the year 2005. Finally, an average reduction in mortality attributable to the decrement of the ambient air particles was quantified analysing the distribution of the percent change between the two years.

4.4.5 Sensitivity analysis

Several analyses were performed in order to study the sensitivity of the results in relationship to: (i) the prior for the DP precision parameter, α, that is the hyperparameter that influences the number of clusters (i.e., mixture components); (ii) the robustness of the results under different initialization of the algorithm (i.e., different initial number of clusters); (ii) the effect of the European heat-wave event in 2003, that London experienced in the period 4 to 13 August 2003 (Johnson et al. 2005; Solberg et al. 2005).

In particular, the sensitivity check was performed as following.

With respect to the prior for α parameter, two different Gamma prior distributions were considered, setting the shape and rate parameters respectively as $a = 2$, $b = 4$ and $a = 1$, $b = 1$.

In term of sampler initialisation, a pretty large number of clusters (that is, $\geq$ 20) was specified to allow the sampler to visit the entire model space (Hastie et al. 2014), and two implementations, with 20 and 30 clusters, were assessed.

Finally, to study the effect of heat wave event, a dummy indicator variable was included in the profile regression model, with value 1 on heat-wave days and 0 otherwise.

4.5 Results

Summary statistics for deaths for respiratory-related diseases and ambient air particles measured in London in the years 2002-2005 are given in Table 4.1.

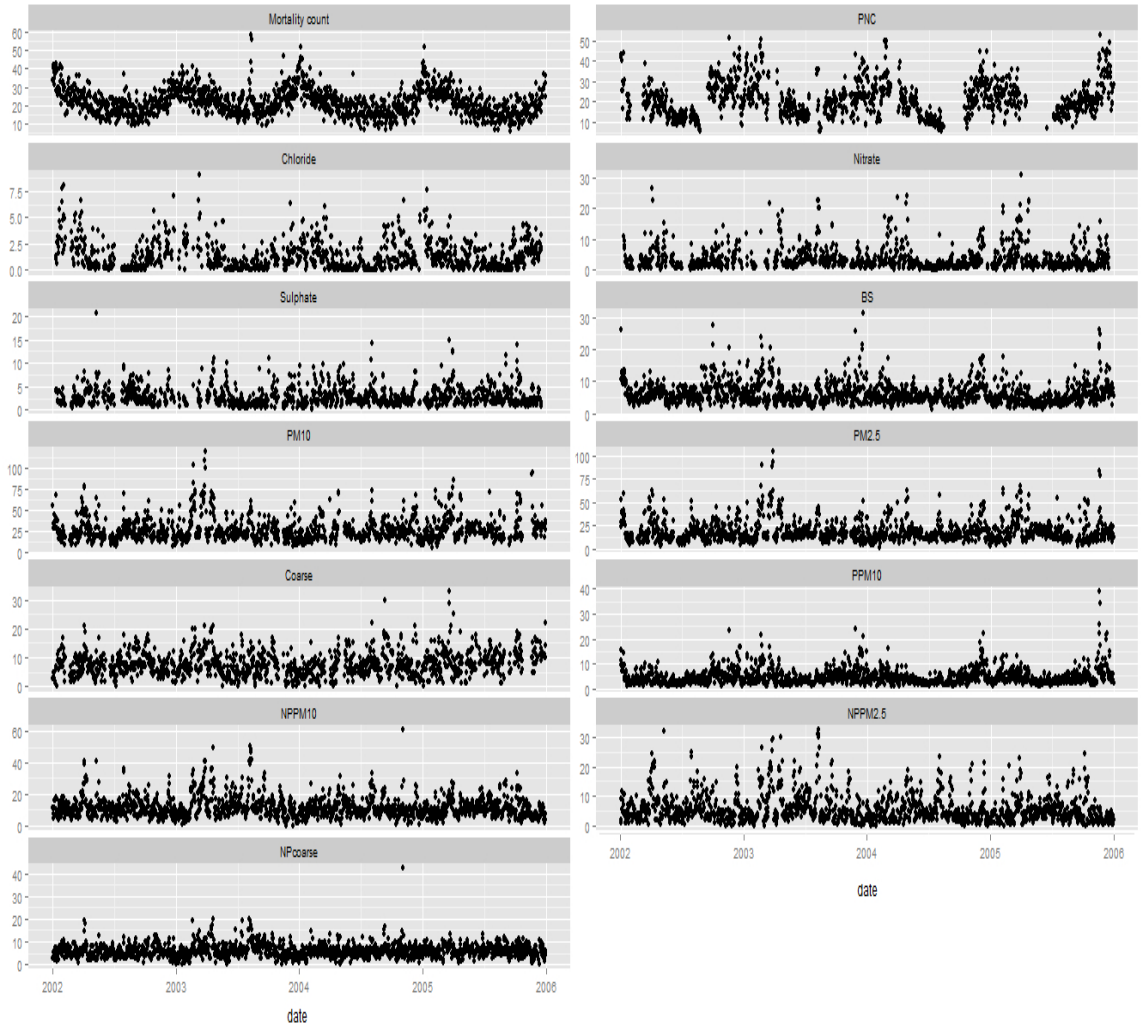
Table 4.1: Descriptive statistics of respiratory mortality and airborne particle metrics. London, 2002-2005.

Variables	Mean	Range	Percentiles 25th	50th	75th
Deaths (per day)	21.60	6.00-58.00	16.00	21.00	26.00
PNC (cm^{-3}/1000)	21.19	5.39-52.44	14.63	19.97	25.91
PM component					
Chloride ($\mu g/m^3$)	1.31	0.01-9.06	0.25	0.88	1.98
Nitrate ($\mu g/m^3$)	3.77	0.03-30.89	1.35	2.44	4.47
Sulphate ($\mu g/m^3$)	2.93	0.23-20.63	1.51	2.25	3.89
BS ($\mu g/m^3$)	6.23	1.40-31.33	4.00	5.40	7.60
PM size					
PM_{10} ($\mu g/m^3$)	26.63	5.00-119.00	17.00	23.00	32.00
$PM_{2.5}$ ($\mu g/m^3$)	18.85	1.00-104.00	11.00	15.00	22.00
Coarse ($\mu g/m^3$)	7.89	0-33.00	5.00	7.00	10.00
PM source apportionment					
PPM_{10} ($\mu g/m^3$)	4.63	0.80-39.10	2.50	3.70	5.60
$NPPM_{10}$ ($\mu g/m^3$)	11.50	0-61.00	7.00	9.90	14.20
$NPPM_{2.5}$ ($\mu g/m^3$)	5.75	0-32.60	2.40	4.20	7.40
NPcoarse ($\mu g/m^3$)	5.99	0-42.20	4.00	5.60	7.40

Figure 4.3 shows the time series of daily mortality counts for respiratory diseases and daily concentrations of airborne particle metrics from London for the years 2002-2005. The data for both mortality and particles exhibited a pronounced seasonal pattern, for example, with mortality increasing during winter months and decreasing during summer months.

The correlations between the daily concentrations of pollutants showed different

Figure 4.3: Daily mortality counts and daily airborne particle metrics in London, 2002-2005.



Particles are plotted in the original measurement units as reported in Table 4.1

.

degrees of interdependence in these metrics, as shown in Table 4.2.

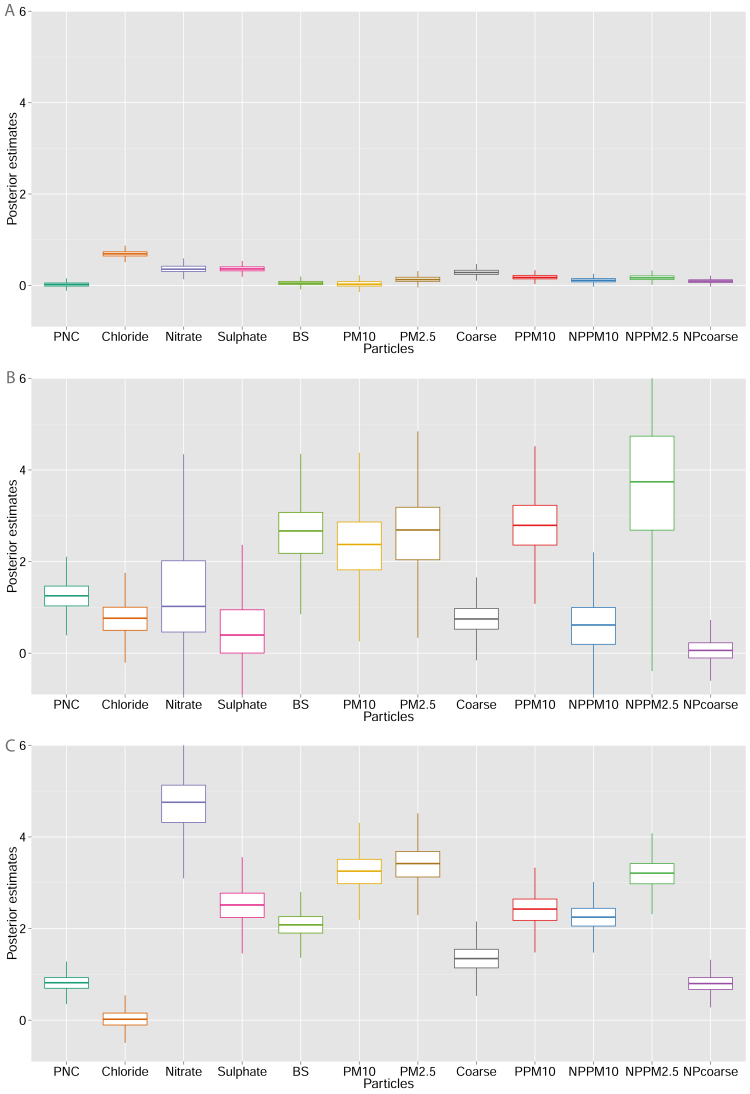
Table 4.2: Correlation between pairs of airborne particle metrics. London, 2002-2005.

	PNC	Chloride	Nitrate	Sulphate	BS	PM_{10}	$PM_{2.5}$	Coarse	PPM_{10}	$NPPM_{10}$	$NPPM_{2.5}$	NPcoarse
PNC	1											
Chloride	0.34	1										
Nitrate	0.38	-0.17	1									
Sulphate	0.08	-0.31	0.52	1								
BS	0.49	-0.16	0.46	0.35	1							
PM_{10}	0.30	-0.16	0.67	0.66	0.48	1						
$PM_{2.5}$	0.31	-0.29	0.70	0.68	0.51	0.91	1					
Coarse	0.09	0.11	0.18	0.25	0.13	0.57	0.26	1				
PPM_{10}	0.72	-0.09	0.53	0.30	0.74	0.53	0.56	0.15	1			
$NPPM_{10}$	-0.12	-0.16	0.43	0.55	0.20	0.68	0.60	0.49	0.11	1		
$NPPM_{2.5}$	-0.16	-0.39	0.48	0.68	0.28	0.67	0.68	0.31	0.14	0.86	1	
NPcoarse	0.02	0.22	0.21	0.15	0.03	0.43	0.25	0.56	0.06	0.71	0.31	1

The representative clustering separated the days into three main clusters, which

included respectively 1156, 63 and 242 days. Figure 4.4 shows the posterior distributions for the particle metrics (on normalised scale) by cluster, while Table 4.3 displays a summary of the cluster multipollutant profiles on their original scale.

Figure 4.4: Box plots showing the distribution of the posterior means for each particle component (on normalised scale) for the three clusters that form the representative clustering (A = cluster 1; B = cluster 2; C = cluster 3).



Compared to clusters 1 and 3, cluster 2 had larger posterior errors as the number of days included was lower.

The risk of mortality for respiratory diseases varied according to these cluster profiles.

Cluster 1 was characterised by low posterior estimates for most of the particles

(except chloride), and had the lowest risk of mortality when compared to the average mortality in 2002-2005. The posterior relative risk of mortality, μ_1, associated with this cluster was 0.98 (95% credible intervals (CI): $0.96, 1.00$).

Cluster 2 was characterised by low posterior estimates of inorganic anions and secondary particles and higher posteriors for primary emissions, with a posterior relative risk of mortality, μ_2, equal to 1.00 (95% CI: 0.97, 1.03). This cluster included mainly winter days.

Finally, cluster 3 was dominated by secondary aerosol, especially nitrate and sulphate, with high posteriors of non-primary airborne particles. The posterior relative risk of mortality, μ_3, was equal to 1.02 (95% CI: $1.00, 1.04$). This third cluster included mainly spring and autumn days.

Table 4.3: Summary of cluster profiles (on original scale): distribution means (95% CI) for characteristics of clusters from the representative clustering.

Particle compounds	cluster 1 (1156 days)	cluster 2 (63 days)	cluster 3 (242 days)
PNC (cm^{-3}/1000)	20.08 (19.54, 20.67)	27.01 (23.63, 30.42)	24.56 (22.58, 26.51)
Chloride ($\mu g/m^3$)	1.38 (1.28, 1.47)	1.43 (0.95, 1.90)	0.90 (0.62, 1.21)
Nitrate ($\mu g/m^3$)	2.90 (2.73, 3.41)	3.76 (2.19, 7.74)	8.58 (6.49, 9.90)
Sulphate ($\mu g/m^3$)	2.61 (2.49, 2.79)	2.65 (1.73, 4.54)	4.76 (3.94, 5.50)
BS ($\mu g/m^3$)	5.48 (5.33, 5.76)	9.80 (7.59, 11.57)	8.83 (7.65, 9.82)
PM_{10} ($\mu g/m^3$)	23.16 (22.51, 25.48)	37.24 (26.94, 45.09)	42.52 (37.61, 47.25)
$PM_{2.5}$ ($\mu g/m^3$)	15.65 (15.12, 17.40)	28.45 (19.10, 35.12)	32.09 (26.84, 35.82)
Coarse ($\mu g/m^3$)	7.57 (7.32, 7.88)	8.87 (7.23, 10.57)	10.36 (8.82, 12.00)
PPM_{10} ($\mu g/m^3$)	3.95 (3.82, 4.22)	7.61 (5.95, 9.70)	7.10 (5.79, 8.06)
$NPPM_{10}$ ($\mu g/m^3$)	10.27 (9.97, 10.73)	11.93 (7.68, 15.86)	17.32 (15.21, 19.46)
$NPPM_{2.5}$ ($\mu g/m^3$)	4.56 (4.34, 5.01)	12.04 (5.41, 18.76)	10.90 (8.74, 12.27)
NPcoarse ($\mu g/m^3$)	5.76 (5.61, 5.91)	5.70 (4.87, 6.63)	6.96 (6.12, 7.86)

Figure 4.5 displays the heatmap of the posterior probabilities that the days (period: 2002-2005) were included in a cluster. For this data set, we found that the days exhibited a high probability of being assigned to a specific cluster.

The posterior estimates for the coefficients associated with the design matrices of B-splines of time and temperature for controlling for seasonal and long-term trend and weather conditions were also analysed. The posterior mean and the 95% CI of the estimated coefficients are displayed in Figure 4.6, showing the effective capability of the model to depict the non-linear effect of these factors.

Figure 4.5: Heatmap of posterior probability that day t belongs to one of the three representative clusters.

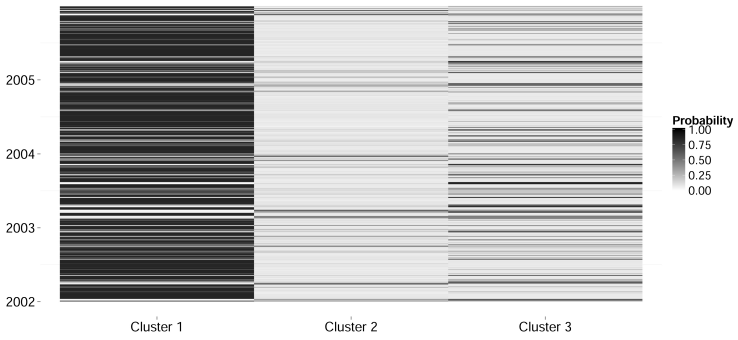
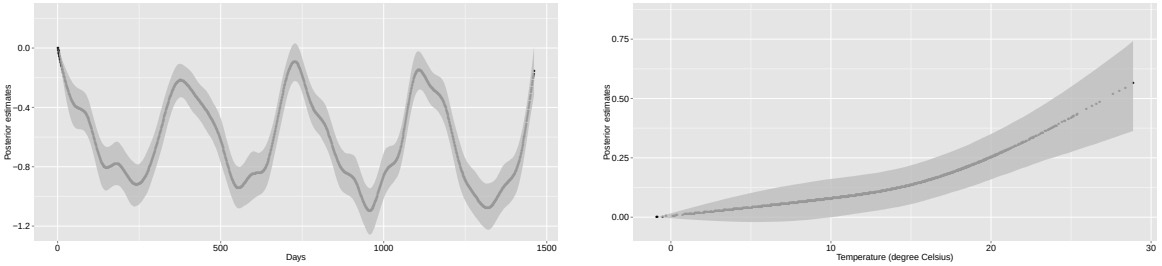


Figure 4.6: Posterior estimates (mean and 95% CI) for the coefficients of the natural cubic spline of time (left panel) and natural cubic spline of temperature (right panel).



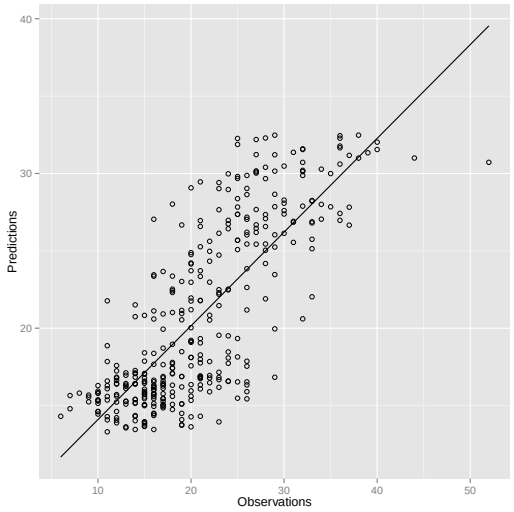
Cross-validation and predictions

Firstly, a cross-validation analysis was performed to check the fit of the model. The respiratory counts of deaths for the year 2005 (here used as validation sample) were predicted using the data 2002-2004 as training sample. The cross-validation produced a R^2 of 0.61 and a RMSE of 8.92. Figure 4.7 provides the scatter plot of validation predictions for the count number of deaths for respiratory diseases in 2005 against the corresponding observations.

Secondly, the combined effect of airborne particles was examined. The predictive distribution of the respiratory mortality counts under the exposure scenario given by the concentrations of particles measured in 2012 was computed. Then this was compared with the predictive distribution obtained for 2005.

Table 4.4 describes the summary statistics for the airborne particles measured in 2012. A large reduction in airborne particles from 2002-2005 to 2012 is clearly visible. This arose mainly from decreases in regional non-primary PM (mainly

Figure 4.7: Scatter plot of validation predictions against observations.



secondary sulphate and nitrate) rather than London specific policies that would have had greater impact on primary PM and BS, consistent with the earlier findings of Fuller and Green (2006). The large decrease in PNC was most likely due to a decrease in the sulphur content of diesel in 2008 which also contributed to decreased sulphate concentrations (Jones et al. 2012).

Table 4.4: Descriptive statistics of airborne particle metrics. London, 2012.

Variables	Mean	Range	Percentiles 25th	50th	75th
PNC (cm^{-3}/1000)	12.12	5.34-25.02	9.16	11.49	14.57
PM component					
Chloride ($\mu g/m^3$)	1.37	0.20-6.40	0.50	1.10	1.80
Nitrate ($\mu g/m^3$)	3.33	0.10-34.40	0.70	1.60	4.00
Sulphate ($\mu g/m^3$)	1.67	0.20-13.50	0.80	1.30	2.10
BS ($\mu g/m^3$)	5.88	1.11-27.78	3.33	4.44	7.41
PM size					
PM_{10} ($\mu g/m^3$)	17.70	4.00-76.00	11.00	14.00	20.75
$PM_{2.5}$ ($\mu g/m^3$)	11.31	2.00-61.00	6.00	8.00	13.00
Coarse ($\mu g/m^3$)	6.60	0-31.00	4.00	6.00	8.00
PM source apportionment					
PPM_{10} ($\mu g/m^3$)	4.11	1.00-14.40	2.30	3.20	5.30
$NPPM_{10}$ ($\mu g/m^3$)	9.49	1.17-29.61	6.12	8.46	11.88
$NPPM_{2.5}$ ($\mu g/m^3$)	3.42	0-17.54	1.35	2.63	4.33
NPcoarse ($\mu g/m^3$)	6.40	0.24-13.47	4.69	6.21	8.00

Comparing the predictive distribution of the deaths for 2012 *vs* 2005, a reduction in respiratory mortality was found, corresponding to an average percentage change in the posterior predictive distributions of -3.51% (95% CI: -0.12%, -5.74%). Based on the observed number of deaths for respiratory-related diseases which occurred in 2005, an average reduction in mortality of approximately 270

subjects would be expected.

Sensitivity analysis

The sensitivity analysis was performed to assess the consistency of the model.

Choices for the prior of the precision parameter α of the DP. The different priors turned out not to have relevant impact on the clustering result. The prior specification of Gamma distribution with: (i) $a = 2$ and $b = 4$ produced a median of 14 clusters, however only three clusters per sweep were well populated (the others included ≤ 9 days); (ii) $a = 1$ and $b = 1$ produced a median of 11 clusters, but again only three clusters per sweep were well populated. The results essentially confirmed the reliability of the three representative clusters obtained in the post-processing.

Different initialization of the algorithm. Setting different starting points in the number of clusters in the initialization of the model, showed the consistency of the results.

Effect of heat wave event in 2003. Using the same seed and the same prior as specified for the main analysis, the results showed that, including this confounding indicator variable, the optimal number of components was reduced to two. In fact, the model resulted in a partitioning solution of the days in groups of 1185 and 276. The smaller cluster of days showed a relative risk of mortality of 1.01 (95%CI: 1.00, 1.02), characterised essentially by high posterior estimates for the most of the metrics, especially secondary particles, with low concentrations of chloride.

4.6 Regression model using temporal profiles of particles from K-means

4.6.1 Clustering of airborne particles

Temporal clustering of airborne particles was performed using K-means partitioning in order to cluster together days with the most similar components and sources, similarly to the DP Bayesian model. As specified in chapter 2 (sec-

tion 2.3.8), K-means algorithm requires three user-specified parameters: (i) the number of clusters K, (ii) cluster initialization, and (iii) the distance metric.

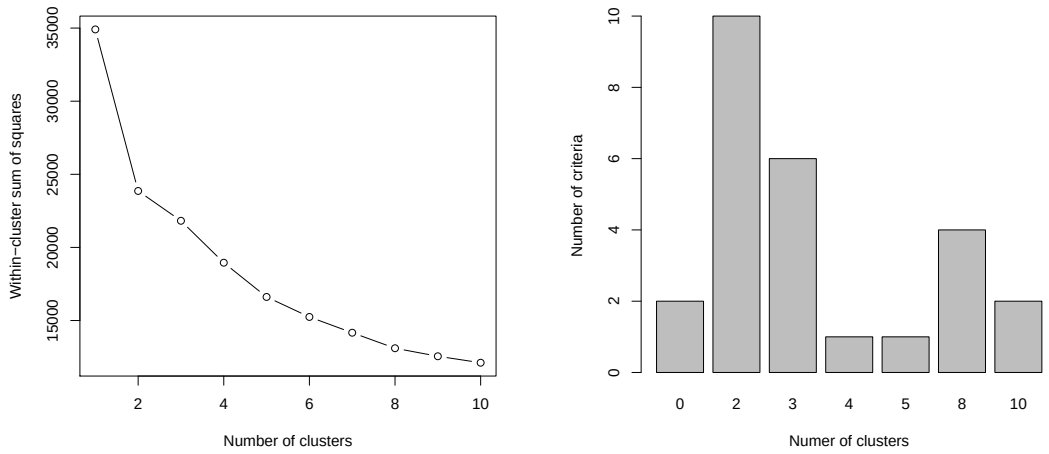
Following Austin et al. (2012), the clustering was performed using the Hartigan and Wong (1979) algorithm, which searches for a local solution that minimizes the Euclidean distance between the observations and the cluster centers. The analysis was performed on normalised data using the function *kmeans* in R (version 3.1.0). Because this algorithm works on complete matrices of data, the set of days available for the analysis was reduced to 667 days of 1461 (45.7%).

The choice of the number of clusters a priori, represents the most critical point in using K-means algorithm. Literature provides several heuristic tools to help in choosing the cluster number, which can be used in addition to pre-existing knowledge of the data or observable features of the data set. An empirical way to find the appropriate number of clusters could be to run K-means clustering with different number of clusters and measure the resulting sum of squared error (SSE) that is the sum of the squared distance between each cluster member and its cluster centroid. A plot of SSE against the number of cluster in K-means solution, could then provide a graphical way to choose an appropriate number of clusters. Specifically, the cluster solution is suggested by the point in which the SSE slows dramatically (that is, the "elbow-point" in the plot). An alternative tool is provided by the R package *NbClust* (version 3.0), that recommends a cluster solution according to the suggestion of 30 criteria. On these data not all the criteria could be calculated, and the output was based on 24 criteria. Figure 4.8 shows the results obtained using these two empirical tools.

Looking to the left panel of Figure 4.8, there is not an obvious break in the distribution of SSE against cluster solutions, however not more than three clusters look be appropriate for this data set. Further, using *NbClust* package, 10 criteria suggested a two clusters solution and six criteria recommended three (right panel).

To validate the cluster solution, the Rand index (Rand 1971) was used. This is an index of external validity and measures the similarity between partition of the same data set. Here was used to compare the two cluster solutions, K =2 or

Figure 4.8: Within-cluster sum of squares for different numbers of clusters (left panel) and suggested number of clusters using the NbClust package (right panel).



K =3. The Rand index assumes values between 0 and 1, where 0 indicates that two data clusters do not agree on any pair of points and 1 indicating that the data clusters are in agreement. Using the R package *clusterCrit* (version 1.2.4) the two K partitions were compared, and the index played a value equal to 1. Thus, the number of K =3 was chosen as final appropriate solution. This was supported also by the DP Bayesian model, performed without pre-specification of the number of cluster, that favored three representative clusters.

Finally, as different initializations of the algorithm can lead to different final clustering (K-means only converges to local minima), it was run with 30 different initial partitions and the partition with the smallest value of the squared error was chosen.

4.6.2 Linking health data and clusters

The approach outlined by Zanobetti et al. (2014) was used to study the association between clustering solution and mortality. In their study, the authors clustered different metrics of chemicals of $PM_{2.5}$ using K-means algorithm and applied a regression-based time series analysis to examine the association of $PM_{2.5}$ with daily total mortality adjusting for long-term trend and seasonality with natural cubic regression splines. Subsequently, Zanobetti and colleagues included an

interaction term between $PM_{2.5}$ and the pollution mixture clusters to assess the effect of different component mixtures.

In the present study, the London data included fewer particle metrics when compared with the study of Zanobetti and colleagues, which comprises rich metrics of PM chemical components. However, with some straining, the analytical strategy of Zanobetti et al. (2014) could be applied. Assuming the Zanobetti's approach, the aim of the analysis was slightly different from the one performed using the DP Bayesian model, and it was focused on exploring the association between PM_{10} with respiratory mortality in London according to different PM_{10} mixtures. Therefore, to allow the inclusion of the interaction term between PM_{10} and cluster within the regression model, the K-means algorithm was re-run excluding PM_{10} from the clustering analysis. The cluster solution was entered in the regression model as categorical variable, and cluster of days with the lowest concentrations in the exposure profile was assumed as reference category (here, it was cluster 3). The model assumed the form:

$$\begin{aligned} &y_t \sim \text{Poisson}(\mu_t), \\ &\log \mu_t = \alpha + \gamma_1 * cl_1 + \gamma_2 * cl_2 + \beta PM_{10} + \delta_1 PM_{10} * cl_1 + \delta_2 PM_{10} * cl_2 + \\ &f_1(t) + f_2(\text{temp}_t) + \epsilon_t \end{aligned} \tag{4.17}$$

were $f_1(t)$ and $f_2(\text{temp}_t)$ represent respectively the spline functions for time and temperature. PM_{10} effect in each cluster was computed by summing β and each δ. As an example, PM_{10} effect in cluster 1 was given by $\beta + \delta_1$ with standard error $\sqrt{var(\beta) + var(\delta_1) + 2cov(\beta, \delta_1)}$.

4.6.3 Results

Table 4.5 presents the clustering solution (on original scale) from the K-means analysis. Cluster 1 included 141 days, characterised by high concentrations of both primary and secondary PM, especially nitrate showed high concentration

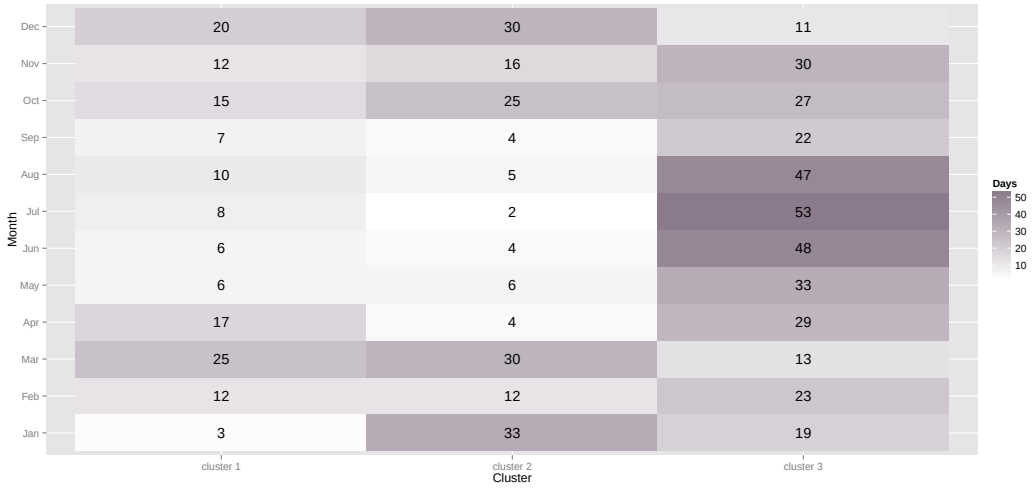
levels. Cluster 2 grouped 171 days and was characterised mainly by elevated concentration of chloride, while Cluster 3, including 355 days, presented low concentrations for all the particle metrics included in the analysis.

Table 4.5: Mean values (standard deviation) of cluster profiles (on original scale) obtained using K-means algorithm.

Particle compounds	cluster 1 (141 days)	cluster 2 (171 days)	cluster 3 (355 days)
PNC (cm^{-3}/1000)	23.84 (9.16)	21.89 (7.43)	16.51 (5.36)
Chloride ($\mu g/m^3$)	0.99 (1.01)	3.04 (1.51)	0.66 (0.62)
Nitrate ($\mu g/m^3$)	8.42 (4.88)	2.27 (1.47)	2.59 (1.87)
Sulphate ($\mu g/m^3$)	5.42 (2.75)	1.99 (1.03)	2.45 (1.29)
BS ($\mu g/m^3$)	9.60 (4.43)	5.29 (2.35)	5.39 (1.89)
PM_{10} ($\mu g/m^3$)	44.84 (12.96)	24.00 (7.12)	20.26 (6.03)
$PM_{2.5}$ ($\mu g/m^3$)	35.08 (11.89)	13.85 (5.69)	14.24 (5.07)
Coarse ($\mu g/m^3$)	9.76 (5.33)	9.35 (3.68)	6.02 (3.01)
PPM_{10} ($\mu g/m^3$)	7.91 (5.38)	4.40 (2.54)	3.71 (1.67)
$NPPM_{10}$ ($\mu g/m^3$)	18.03 (5.39)	10.46 (3.80)	8.87 (4.19)
$NPPM_{2.5}$ ($\mu g/m^3$)	10.69 (6.06)	3.00 (2.25)	4.31 (3.15)
NPcoarse ($\mu g/m^3$)	7.34 (4.41)	7.46 (2.22)	4.55 (2.01)

To assess the seasonal distribution of days according cluster, a heatmap of the daily particle profiles aggregated by month was performed (Figure 4.9). From this analysis, it was clear that (i) Cluster 1 was mainly constituted by spring and autumn days, (ii) Cluster 2 occurred more frequently in winter period, and (iii) Cluster 3 included most of the summer days.

Figure 4.9: Heatmap of cluster frequency by month



The results from the time series regression analysis performed to assess the mortality effect of PM_{10} according its differential composition in mixtures (as detected by K-means clustering) are presented in table 4.6.

Table 4.6: Percent increase (95% confidence intervals), in respiratory mortality for 10 $\mu g/m^3$ increase in PM_{10} and specific cluster effect (reference category: Cluster 3).

Parameters	RR (95% confidence intervals)
Total PM_{10}	1.03 (0.98, 1.07)
Cluster 1	1.01 (0.98, 1.05)
Cluster 2	1.00 (0.96, 1.03)

According this analysis there was no evidence of an effect of particle metrics on respiratory mortality. However, it is important to underline that these results were based on a quite reduced data set (including minus of the 50% of original data), because of missing measurements in the exposure metrics, and this element produced a low statistical power to detect a possible PM effect.

4.7 Discussion

There is an increasing need to assess the health effects of multiple air pollution exposures for both health research and air quality management. This requires new statistical methods to better understand these complex systems.

This study addressed the problem by introducing a Bayesian modelling framework that offers a flexible way to model the joint distribution of a response and pollutants. The proposed model is based on the DP mixture models that represent an appealing tool for clustering data. In standard applications, however, these models assume that the observations are exchangeable and the data points do not have an inherent order influencing their labelling. Several Bayesian nonparametric studies have been specially targeted to clustering temporally evolving phenomena. For example, in a recent work Nieto-Barajas and Contreras-Cristán (2014) accommodated the temporal effects in time series data using a first order autoregressive process. In this model, a simple and feasible solution was used, given by introducing natural cubic splines that correct for temporally dependent confounding effects, adjusting for seasonal and long-term trends and weather variables such as temperature.

A clear benefit of the Bayesian model proposed is the simultaneous estimation

of the contribution of all pollutants to the mortality risk. This would allow policy makers to have a holistic picture of the effect of complex air pollution mixtures. This is a novel feature of the model, in comparison to the recent two-stage approaches proposed by Matyasovszky et al. (2011) and Zanobetti et al. (2014) for example. In this way, in fact, the outcome of interest influences the cluster membership. The model, moreover, presents additional advantages compared with traditional clustering methods such as K-means. First, it is able to address the challenging question of uncertainty in the cluster assignment. In the presented application was found that the uncertainty associated with the partitioning of the days to clusters was quite low, and this supports the use of the partitioning around medoids method on the posterior dissimilarity matrix to obtain a representative partition. Once this partition was obtained, full uncertainty about its characteristics was recovered from post processing of the full MCMC output. Second, because of the Bayesian computation method adopted, the whole time series of particles were considered, without the exclusion of days with missing measurements. Using K-means algorithm only the $45,7\%$ of the days could be included in the analysis. Using the Bayesian model, missing values in a (daily) covariate profile were sampled within the MCMC sampler (i.e., it checked which cluster the day was allocated to and then sampled). Finally, the model was able to uncover clusters in the data naturally, without a clustering structure being imposed by the user.

However, compared to non-Bayesian methods, the model had higher computational cost. K-means algorithm converged quickly, while the Bayesian model took approximately 25 min using an Intel(R) Core i7 CPU machine (2.40GHz, 8 GB RAM) for the inferential procedures. This sacrifice in terms of computational effort is, however, reasonable given the advantages provided by the Bayesian approach with MCMC inference.

The model was applied to a real data set, in which the temporal structure of particle mix in London and its effect on respiratory mortality was studied. It identified which type of pollutant mixtures were associated with mortality and

quantified the risk; in this case a relative risk of mortality was 1.02 (95% CI: 1.00, 1.04) on days with increased secondary PM mass concentrations (i.e., metrics not associated with NO_x sources) including high concentrations of inorganic PM such as sulphate and nitrate.

The previous study of Atkinson et al. (2010) found, however, association between respiratory mortality and particle mass concentrations that could not be explained by sulphate and nitrate at all lags. Compared with the results from Atkinson et al. (2010), this finding is more consistent with the large contribution of sulphate and nitrate to PM mass concentrations.

Particulate nitrate and sulphate are acidic in nature. Nitrate is mainly the product of oxidation of nitrogen oxides (which sources include fossil-fuel combustion; road transport, space heating and aircraft for example, biomass burning, soil release and ammonia oxidation from agriculture), while sulphate is mainly from the oxidation of sulphur dioxide (emitted from power plants and industrial facilities and to a lesser extent natural sources such as oceans, plant and soils, and volcanoes along with ammonia oxidation). Evidence of associations between secondary inorganic PM, such as sulphates and nitrates with negative health effects are limited and still insufficient to support a causality (Reiss et al. 2007; WHO/Europe 2013). However, the results of our study for respiratory mortality are consistent with Ostro et al. (2009), which observed an increased risk of respiratory hospital admissions in children associated with an increase in sulphate for a 3-day lag. Recently, Dai et al. (2014) found that particle sulphur modified the effect of $PM_{2.5}$ on total and respiratory mortality. As sulphate is the primary form of particle sulphur, the authors interestingly argued about the plausibility of the health effects of sulphate, supported by toxicology findings that show, for example, that it is linked to an increased oxidative stress and coagulation (Chuang et al. 2007). Cao et al. (2012) found significant positive associations of total, cardiovascular, and respiratory mortality with different PM components, including nitrate, at 1 day lag.

Rather than producing single-pollutant concentration response functions for

use in health impact assessment or to assess the cost benefits of policies to decrease pollution exposures, the Bayesian approach provides a predictive tool to allow the assessment of changes in the pollutant mixture. This is a far more realistic representation of the outcomes of the range of policies being employed across different emissions sectors at different spatial and government levels rather than taking a single pollutant approach. When assessing impact through a single pollutant approach, it is unclear if the concentration response function for a single pollutant is acting as a tracer for health effects from other correlated pollutants; for instance Janssen et al. (2012) have examined if black carbon particles or PM mass concentrations are a better metric for airborne particle health effects. These issues are avoided by instead looking at mixtures. As an illustration of this approach we estimated the changes in health response from changes in pollution concentrations in all 12 exposure variables measured in our data set. Between 2005 and 2012, a decrease in annual respiratory mortality of -3.51% (95% CI: -0.12%, -5.74%) in London was estimated.

This study has several limitations. It was ecological and the measurements of particle metrics were collected at a single monitoring site in central London, therefore we could not account for individual features and activities. It is commonly accepted that in population-based time series studies, individual risk factors (age, diet, smoking etc.) are unlikely to be confounders as they do not vary temporally with air pollution over relatively short-term periods (e.g., Burnett et al. 2003; Sheppard et al. 2012). However, the ambient measurements used in our study could lack of spatial and temporal resolution due to individual's activities (Özkaynak et al. 2013) and generally be less representative than personal monitoring for assessing particulate exposure (Buonanno et al. 2013). Moreover, respiratory mortality in London population was related only to outdoor particle concentrations, while people spend considerable time in indoor environments and exposure highly depends on indoor concentrations (Morawska et al. 2013).

At the time of the original study of Atkinson et al. (2010), only limited information on PM composition were available. A more in-depth understanding

of the dynamics in pollution mixtures will be provided by the inclusion of more detailed chemical speciation of PM. Finally, the Bayesian model proposed in this work has only considered daily mortality from respiratory causes but it could equally be applied to other outcomes, namely daily cardiovascular mortality and cardiorespiratory hospital admissions.

5 | Conclusions and future work

5.1 Concluding remarks

This thesis has been motivated by the need to investigate statistical methodologies for characterising exposure metrics of particle components and sources, to be used in short-term health effect studies.

Unlike gaseous pollutants, such as O_3, SO_2, NO_2, or CO, airborne particles comprise a complex mixture of both primary and secondary compounds, with different chemical and physical features, and associations between particles and health outcomes in epidemiologic studies may be the result of multiple components and/or sources acting on different physiological mechanisms. The thesis was built up assuming this mixture view of particle metrics.

I addressed the topic from a statistical standpoint, with specific emphasis on the construction and characterisation of exposure modelling.

Chapter 1 provided the rational for the thesis along with health and policy context and defined specific aims.

Chapter 2 summarised the key issues faced, from a statistical view point, in analysing air pollution exposure metrics and health outcomes arranged in time series, and synthesised methods gathered from various approaches in a logical manner. Although reviews exist on this subject (e.g., Pitard and Viel 1997; Dominici et al. 2010; Billionnet et al. 2012; Sun et al. 2013; Oakes et al. 2014), most of them have provided a thorough treatment of several methods, omitting others (especially if developed under a Bayesian paradigm), and in several cases the framework adopted ranged from cross-sectional studies to toxicological ana-

lyses. Thus, chapter 2 was devoted to provide an overview of methods organising and discussing them in a organic way, showing their pros and cons.

One of the main drawback of the classical methods for air pollution time series studies, is the lack of the spatial dimension in the characterisation of exposure models. Indeed, in most of the studies, the geographical domain is assumed well represented by only one or a few monitoring stations, expressing the average concentrations experienced by a community. Exposure metrics of air pollution are, however, complex, especially when arising from large urban environment, where components from local sources, mainly traffic-related, mix with secondary or natural compounds made up of PM formed from gaseous precursors. Chapter 3 proposed a full Bayesian hierarchical approach to this problem within a data assimilation perspective, combining ambient measurements from ground-based monitoring stations located in urban an rural areas and output from a dispersion air quality model capturing the local-scale component of PM using a sophisticated description of the relevant physical and chemical processes taking place in atmosphere, as well as space and time varying covariates. A number of models, highly flexible in structure thanks to time- and space-varying coefficients, were compared, showing different degree of predictive ability. From this study, it clearly emerged that the use of the local-scale air pollution simulations to describe urban pollution levels should be combined with regional background sources and that estimates of daily pollution concentrations can be improved by including space-time varying factors to account for residual variations that are not present in emission inventories, such as day of the week (as emissions inventories only give annual totals) and temperature (to describe seasonal changes in chemistry between primary and regional secondary PM). The inclusion of these different information concerning the geographical domain produced a good performance of the proposed structures in comparison to standard space-time statistical modelling approaches which typically present varying intercepts. Indeed, the latter could be appropriate in research situations where relatively poor source inform-

ation are available, thus the random effects can be thought as latent variables which capture the effects of unknown or unmeasured space-time covariates. Finally, the Bayesian hierarchical framework adopted could be used in future air pollution exposure modelling, as it (i) offered a natural way for combining different particle sources, and properly accounted for their associated uncertainties, and (ii) provided a way to include previous scientific knowledge about temporal dependency and spatial correlation.

Chapter 4 provided a flexible Bayesian semiparametric model for clustering time points with similar particle and health response profiles, with the aim of distinguishing the potential harmful effects of exposure metrics constituted by components with different chemical and physical features, and characterised by different degree of correlation. In the application on London time series of particles and respiratory mortality, cluster membership seemed to be an effect modifier in health effects analysis, denoting pollutant mixtures that could be targeted as part of air quality control strategy for health. In particular, it showed higher risk of respiratory mortality associated with spring secondary pollution episodes.

In recent years, nonparametric techniques have been widely employed for clustering data, where the number of clusters is inferred in a data-driven way. To the best of my knowledge, no specific contributions have been performed, within this paradigm, for clustering time series of pollutants and health responses simultaneously. Often nonparametric clustering methods do not explicitly exploit the order information contained in the data, and assume the observations are exchangeable. Although this assumption usually leads to a easy tractable model form for the posterior computation, it may degrade the clustering performance as the temporal order of the data is not accounted for. To address this issue, the model resolved with the inclusion of flexible spline functions to control for temporal trend and seasonal variability. Furthermore, the comparison of the proposed modelling approach with benchmarked clustering methods such as K-means, allowed the identification a number of attractive advantages. Unlike traditional model-free clustering methods in which clusters are formed on the basis of inter-

cluster distances, with no concept of sequentiality in the data, Bayesian profile regression equipped with spline functions grouped time points on the basis of probability estimated from mixture modelling and properly accounted for the uncertainty associated with the clustering procedure rather than merely giving a single partition solution. Moreover, this Bayesian model incorporated the association with the health outcome in determining the particle profile that characterised cluster membership. Finally, it showed the capability of Bayesian methods in dealing with the notable issue of missing data in the air pollution exposure metrics. This approach, also, could provide a new technique for policy makers to assess the impact of interventions that affect the mixture rather than individual pollutants. This reflects the reality of air pollution management strategies. For instance, the progressive restrictions on vehicle emission through euro-standards have acted on several pollutant simultaneously.

In the light of my work, I would conclude that a multiple pollutant and source approach represents a more natural and effective way to deal with the air polluted problem. Estimating the health effects deriving from exposure to a single pollutant or a source is a useful analytical construct, however it is not representative of true exposure. People are actually exposed to mixtures, not to a single pollutant at a time, and how mixtures of pollutants affect health represents now a challenging goal to advance the knowledge on the relationships between pollution and health.

5.2 Future work

The work described previously leaves some questions open in modelling particle metrics that can be further investigated. In particular:

- The spatio-temporal Bayesian models presented in chapter 3 included the spatial process through a Bayesian treatment of Kriging in which the covariance model, used in the Gaussian process prior distribution for the spatial

field, was stationary. Although realistic for an urban environment such as London, in other geographical domains, where the spatial covariance of atmospherically driven pollutants could be affected by spatially varying features of landscape, topography and interaction of meteorology and emissions, this form of covariance could be too simple. Therefore, non-stationary covariance structures might be explored and the models might be compared in their predictive abilities.

- The Bayesian profile regression presented in chapter 4 included natural cubic splines to correct for temporally dependent confounding effects, adjusting for seasonal and long-term trends and weather variables such as temperature. A possible alternative might be given by considering a probit stick-breaking prior for the weights of the process, as suggested by Rodríguez and Dunson (2011). In particular, the authors proposed a new construction for the weights obtained by replacing the characteristic Beta distribution in the definition of the sticks by probit transformations of normal random variables. Rodríguez and Dunson (2011) showed that this new class of priors is able to capture the time-evolving statistical properties of time series data in a finance context. Thus, might be appealing explore this prior scheme in air pollution time series studies.

  Furthermore, Bayesian profile regression seemed a promising tool to be used for health impact scenario assessment. This, however, would require the testing of this methodology in other urban environments, as well as the inclusion of a richer chemical speciation of PM (for example, including in the analysis organic and elemental carbon along with metal species and oxidative potential).

- The spatio-temporal hierarchical modelling approach in chapter 3 proposed structures for combining different particle data sources; the semiparametric profile regression in chapter 4 described a joint model for mixture of particles and health response; a third model concerning with the use of particle

mixtures to identify sources is, at the moment, under development.

The aim of this study would be to develop a new Bayesian multivariate receptor model for estimating source contributions to PM, accounting for meteorological factors such as wind speed and wind direction (translated on Cartesian coordinates) and capturing dynamics in the sequence of data. The model would be compared to three different methods for source apportionment of atmospheric aerosol, specifically: the benchmarked methods of positive matrix factorization and K-means partitioning, and the Bayesian profile regression model. The data available for this study consist of concentrations of 27 chemical components of PM_{10} collected at Mülheim-Styrum site in North Rhine-Westphalia, Germany, between April 2008 and March 2009. Data on wind speed and direction are available at the same site and for the same period. Results from the three comparative models have already been obtained. Finally, a synthetic data set to validate the proposed modelling approach has been generated, simulating time series data from a multivariate normal distribution that depend on an autoregressive model of first order, while accounting for a specific indicator of seasonal window. To simulate these particle data, estimates about the variance-covariance matrix were obtained from monitored chemical component concentrations at North Kensington background site in London. The new Bayesian model is under development using the software WinBUGS interfaced with R software.

Bibliography

ACGIH (1994), *1994-1995 Threshold limit values for chemical 16 substances and physical agents and biological exposure indices*, Cincinnati, OH: American Conference of Governmental Industrial Hygienists.

Aguilar, O., Huerta, G., Prado, R., and West, M. (1998), "Bayesian inference on latent structure in time series," *Bayesian Statistics*, 6, 1–16.

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium on Information Theory*, eds. Petrov, B. N. and Csáki, F., pp. 267–281.

Allen, G. and Reiss, R. (1997), "Evaluation of the TEOM method for measurement of ambient particulate mass in urban areas," *The Journal of the Air & Waste Management Association*, 47, 682–689.

Anenberg, S. C., Horowitz, L. W., Tong, D. Q., and West, J. J. (2010), "An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling," *Environmental Health Perspectives*, 118, 1189–1195.

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, 2, 1152–1174.

Armstrong, B. G. (1998), "Effect of measurement error on epidemiological studies of environmental and occupational exposures," *Occupational and Environmental Medicine*, 55, 651–656.

Assunção, R. M. (2003), "Space varying coefficient models for small area data," *Environmetrics*, 14, 453–473.

Atkinson, R. W., Fuller, G. W., Anderson, R. H., Harrison, R. M., and Armstrong, B. (2010), "Urban ambient particle metrics and health: a time-series analysis," *Epidemiology*, 21, 501–511.

Atkinson, R. W., Kang, S., Anderson, R. H., Mills, I. C., and Walton, H. A. (2014), "Epidemiological time series studies of $PM_{2.5}$ and daily mortality and hospital admissions: a systematic review and meta-analysis," *Thorax*, 69, 660–665.

Austin, E., Coull, B., Thomas, D., and Koutrakis, P. (2012), "A framework for identifying distinct multipollutant profiles in air pollution data," *Environment International*, 45, 112–121.

Bagheri, A., Midi, H., and Imon, A. H. M. R. (2010), "The effect of collinearity-influential observations on collinear data set: a Monte Carlo simulation study," *Journal of Applied Sciences*, 10, 2086–2093.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by supervised principal components," *Journal of the American Statistical Association*, 101, 119–137.

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall/CRC, 2nd ed.

Baron, P. A. and Willeke, K. (2001), *Aerosol measurement: principles, techniques, and applications*, John Wiley & Sons, 2nd ed.

Bateson, T. F., Coull, B. A., Hubbell, B., Ito, K., Jerrett, M., Lumley, T., Thomas, D., Vedal, S., and Ross, M. (2007), "Panel discussion review: session 3 - Issues involved in interpretation of epidemiologic analyses - statistical modeling," *Journal of Exposure Science and Environmental Epidemiology*, 17, S90–S96.

Bateson, T. F. and Wright, J. M. (2010), "Regression calibration for classical exposure measurement error in environmental epidemiology studies using multiple local surrogate exposures," *American Journal of Epidemiology*, 172, 344–352.

Belis, C. A., Karagulian, F., Larsen, B. R., and Hopke, P. K. (2013), "Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe," *Atmospheric Environment*, 69, 94–108.

Bell, M. L., Ebisu, K., Leaderer, B. P., Gent, J. F., Lee, H. J., Koutrakis, P., Wang, Y., Dominici, F., and Peng, R. D. (2014), "Associations of $PM_{2.5}$ constituents and sources with hospital admissions: analysis of four counties in Connecticut and Massachusetts (USA) for persons ≥ 65 years of age," *Environmental Health Perspectives*, 122, 138–144.

Bell, M. L., Ebisu, K., and Peng, R. D. (2011), "Community-level spatial heterogeneity of chemical constituent levels of fine particulates and implications for epidemiological research," *Journal of Exposure Science and Environmental Epidemiology*, 21, 372–384.

Bell, M. L., Ebisu, K., Peng, R. D., Samet, J. M., and Dominici, F. (2009), "Hospital admissions and chemical composition of fine particle air pollution," *American Journal of Respiratory and Critical Care Medicine*, 179, 1115–1120.

Bell, M. L., Samet, J. M., and Dominici, F. (2004), "Time-series studies of particulate matter," *Annual Review of Public Health*, 25, 247–280.

Bell, M. L., Zanobetti, A., and Dominici, F. (2013), "Evidence on vulnerability and susceptibility to health risks associated with short-term exposure to particulate matter: A systematic review and meta-analysis," *American Journal of Epidemiology*, 178, 865–876.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons, Inc.

Benjamini, Y. and Yekutieli, D. (2005), "False discovery rate-adjusted multiple confidence intervals for selected parameters," *Journal of the American Statistical Association*, 11, 71–81.

Berliner, L. M. (1996), "Hierarchical Bayesian time series models," in *Maximum Entropy and Bayesian Methods*, eds. Hanson, K. and Silver, R., Dordrecht, NL: Kluwer Academic Publishers, pp. 15–22.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010a), "A bivariate space-time downscaler under space and time misalignment," *Annals of Applied Statistics*, 4, 1942–1975.

— (2010b), "A spatio-temporal downscaler for output from numerical models," *Journal of Agricultural, Biological and Environmental Statistics*, 15, 176–197.

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L., and Armstrong, B. (2013), "Time series regression studies in environmental epidemiology," *International Journal of Epidemiology*, 42, 1187–1195.

Biggeri, A., Bellini, P., and Terracini, B. (2004), "Meta-analysis of the Italian studies on short-term effects of air pollution - MISA 1996-2002," *Epidemiologia e prevenzione*, 28(4-5 Suppl), 4–100.

Billionnet, C., Sherrill, D., Annesi-Maesano, I., and GERIE study (2012), "Estimating the health effects of exposure to multi-pollutant mixture," *Annals of Epidemiology*, 22, 126–141.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, New York: Springer-Verlag.

Blackwell, D. (1973), "Discreteness of Ferguson Selections," *Annals of Statistics*, 1, 356–358.

Blackwell, D. and MacQueen, J. B. (1973), "Ferguson distributions via Polya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

Blangiardo, M., Richardson, S., Gulliver, J., and Hansell, A. (2011), "A Bayesian analysis of the impact of air pollution episodes on cardio-respiratory hospital admissions in the Greater London area," *Statistical Methods in Medical Research*, 20, 69–80.

Bobb, J. F., Dominici, F., and Peng, R. D. (2013), "Reduced Bayesian hierarchical models: estimating health effects of simultaneous exposure to multiple pollutants," *Journal of the Royal Statistical Society: Series C*, 62.

Bressi, M., Sciare, J., Ghersi, V., Bonnaire, N., Nicolas, J. B., Petit, J.-E., Moukhtar, S., Rosso, A., Mihalopoulos, N., and Féron, A. (2013), "A one-year comprehensive chemical characterisation of fine aerosol ($PM_{2.5}$) at urban, suburban and rural background sites in the region of Paris (France)," *Atmospheric Chemistry and Physics*, 13, 7825–7844.

Briggs, D. J., de Hoogh, C., Guiliver, J., Wills, J., Elliott, P., Kingham, S., and Smallbone, K. (2000), "A regression-based method for mapping traffic related air pollution: application and testing in four contrasting urban environments," *Science of The Total Environment*, 253, 151–167.

Brown, J. S., Gordon, T., Price, O., and Asgharian, B. (2013), "Thoracic and respirable particle definitions for human health risk assessment," *Particle and Fibre Toxicology*, 10, 12.

Bruno, F. and Cocchi, D. (2002), "A unified strategy for building simple air quality indices," *Environmetrics*, 13, 243–261.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear smoothers and additive models," *The Annals of Statistics*, 17, 453–510.

Buonanno, G., Marks, G. B., and Morawska, L. (2013), "Health effects of daily airborne particle dose in children: direct association between personal dose and respiratory health effects," *Environmental Pollution*, 180, 246–250.

Burnett, R. T., Brook, J., Dann, T., Delocla, C., Philips, O., Cakmak, S., Vincent, R., Goldberg, M. S., and Krewski, D. (2000), "Association between particulate and gas phase components of urban air pollution and daily mortality in eight Canadian cities," *Inhalation Toxicology*, 12, 15–39.

Burnett, R. T., Dewanji, A., Dominici, F., Goldberg, M. S., Cohen, A., and Krewski, D. (2003), "On the relationship between time-series studies, dynamic population studies, and estimating loss of life due to short-term exposure to environmental risks," *Environmental Health Perspectives*, 111, 1170–1174.

Cameletti, M., Ignaccolo, R., and Bande, S. (2011), "Comparing spatio-temporal models for particulate matter in Piemonte," *Environmetrics*, 22, 985–996.

Cao, J., Xu, H., Xu, Q., Chen, B., and Kan, H. (2012), "Fine particulate matter constituents and cardiopulmonary mortality in a heavily polluted Chinese city," *Environmental Health Perspectives*, 120, 373–378.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, London: Chapman & Hall Ltd, 2nd ed.

Carruthers, D. J., Edmunds, H. A., Lester, A. E., McHugh, C. A., and Singles, R. J. (2000), "Use and validation of ADMS-Urban in contrasting urban and industrial locations," *International Journal of Environment and Pollution*, 14, 364–374.

Carslaw, D. C. and Ropkins, K. (2012), "openair - An R package for air quality data analysis," *Environmental Modelling & Software*, 27-28, 52–61.

Chang, H. H., Peng, R. D., and Dominici, F. (2011), "Estimating the acute health effects of coarse particulate matter accounting for exposure measurement error," *Biostatistics*, 12, 637–652.

Chatfield, C. (2004), *The Analysis of Time Series. An Introduction*, London: Chapman & Hall, 6th ed.

Chatterjee, S., Hadi, A. S., and Dunn, G. (2000), *Regression Analysis by Examples*, New York: Wiley, 2nd ed.

Chiogna, M. and Gaetan, C. (2002), "Dynamic generalized linear models with applications to environmental epidemiology," *Journal of the Royal Statistical Society: Series C*, 51, 453–468.

Choi, J., Fuentes, M., and Reich, B. J. (2009), "Spatial-temporal association between fine particulate matter and daily mortality," *Computational Statistics and Data Analysis*, 53, 2989–3000.

Chow, J. C. and Watson, J. G. (2002), "Review of $PM_{2.5}$ and PM_{10} apportionment for fossil fuel combustion and other sources by the chemical mass balance receptor model," *Energy & Fuels*, 16, 222–260.

Christensen, W. F. and Sain, S. R. (2002), "Accounting for dependence in a flexible multivariate receptor model," *Technometrics*, 44, 328–337.

Christensen, W. F., Schauer, J. J., and Lingwall, J. W. (2006), "Iterated confirmatory factor analysis for pollution source apportionment," *Environmetrics*, 17, 663–681.

Chuang, K. J., Chan, C. C., Su, T. C., Lee, C. T., and Tang, C. S. (2007), "The effect of urban air pollution on inflammation, oxidative stress, coagulation, and autonomic dysfunction in young adults," *American Journal of Respiratory and Critical Care Medicine*, 176, 370–376.

Chuang, Y.-H., Mazumdar, S., Park, T., Tang, G., Arena, V. C., and Nicolich, M. J. (2010), "Bayesian model averaging approach in health effects studies: Sensitivity analyses using PM_{10} and cardiopulmonary hospital admissions in Allegheny County, Pennsylvania and simulated data," *Atmospheric Pollution Research*, 1, 161–167.

City Mayors Statistics (2012), "Europe's largest cities - Cities ranked 1 to

100," http://www.citymayors.com/features/euro_cities1.html, last accessed date 25 June 2014.

Clyde, M. (2000), "Model uncertainty and health effect studies for particulate matter," *Environmetrics*, 11, 745–763.

Cocchi, D., Greco, F., and Trivisano, C. (2007), "Hierarchical space-time modelling of PM_{10} pollution," *Atmospheric Environment*, 41, 532–542.

Cox, L. H. (2000), "Statistical issues in the study of air pollution involving airborne particulate matter," *Environmetrics*, 11, 611–626.

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005), "Bayesian analysis for penalized spline regression using WinBUGS," *Journal of Statistical Software*, 14, 1–24.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley.

Cressie, N. and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, New York: Wiley.

Dai, L., Zanobetti, A., Koutrakis, P., and Schwartz, J. D. (2014), "Associations of fine particulate matter species with mortality in the United States: A multicity time-series analysis," *Environmental Health Perspectives*, 122, 837–842.

Daniels, M. J., Dominici, F., and Zeger, S. (2004), "Underestimation of standard errors in multi-site time series studies," *Epidemiology*, 15, 57–62.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.

Deligiorgi, D. and Philippopoulos, K. (2011), "Spatial interpolation methodologies in urban air pollution modeling: application for the Greater Area of Metropolitan Athens, Greece," in *Advanced Air Pollution*, ed. Nejadkoorki, F., ISBN: 978-953-307-511-2, InTech, DOI: 10.5772/17734.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1997), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, 39, 1–38.

Denby, B., andDavid Holland, V. G., and Hogrefe, C. (2009), "Integration of air quality modeling and monitoring data for enhanced health exposure assessment," *EM: Air and Waste Management Associations Magazine for Environmental Managers. Air & Waste Management Association*, Pittsburgh, PA, (10/2009), 46–49.

Department of Health (1998), *Committee on the medical effects of air pollutants. Quantification of the effects of air pollution on health in the United Kingdom*, London, UK: The Stationery Office.

DETR (1999), "Assistance with the Review and Assessment of PM_{10} Concentration in Relation to the Proposed EU Stage 1 limit values," .

Diggle, P. and Ribeiro, P. J. (2007), *Model-Based Geostatistics*, New York: Springer.

Dominici, F. (2004), *Time-series analysis of air pollution and mortality: a statistical review*, Cambridge, MA: Health Effects Institute, Research Report 123.

Dominici, F., McDermott, A., and Hastie, T. J. (2004), "Improved semiparametric time series models of air pollution and mortality," *Journal of the American Statistical Association*, 99, 938–948.

Dominici, F., Peng, R. D., Barr, C. D., and Bell, M. L. (2010), "Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach," *Epidemiology*, 21, 187–194.

Dominici, F., Samet, J. M., and Zeger, S. L. (2000), "Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy," *Journal of the Royal Statistical Society: Series A*, 163, 263–302.

Dominici, F., Sheppard, L., and Clyde, M. (2003), "Health effects of air pollution: a statistical review," *International Statistical Review*, 71, 243–276.

Dominici, F., Wang, C., Crainiceanu, C., and Parmigiani, G. (2008), "Model selection and health effect estimation in environmental epidemiology," *Epidemiology*, 18, 558–560.

Draper, D. (1995), "Assessment and propagation of model uncertainty," *Journal of the Royal Statistical Society: Series B*, 57, 45–97.

Dunson, D. B., Pillai, N., and Park, J. H. (2007), "Bayesian density regression," *Journal of the Royal Statistical Society: Series B*, 69, 163–183.

Eilers, P. H. C. and Marx, B. D. (1996), "Flexible smoothing using B-splines and penalties (with discussion)," *Statistical Science*, 11, 89–121.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Everitt, B. S. and Dunn, G. (2001), *Applied Multivariate Data Analysis*, London: Edward Arnold, 2nd ed.

Fahrmeir, L. and Lang, S. (2001), "Bayesian inference for generalized additive mixed models based on Markov random field priors," *Applied Statistics*, 50, 201–220.

Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer.

Fan, J. and Zhang, W. (2008), "Statistical methods with varying coefficient models," *Statistics and Its Interface*, 1, 179–195.

Ferguson, T. S. (1973), "A Bayesian analysis of some non-parametric problems," *The Annals of Statistics*, 1, 209–230.

— (1974), "Prior distributions on spaces of probability measures," *The Annals of Statistics*, 2, 615–629.

Finlayson-Pitts, B. J. and Pitts, J. N. (2000), *Chemistry of the Upper and Lower Atmosphere. Theory, Experiments, and Applications*, San Diego, CA: Academic Press.

Fox, E. B. and Jordan, M. I. (2013), "Mixed membership models for time series," in *Handbook on Mixed Membership Models*, eds. Airoldi, E., Blei, D., Erosheva, E., Fienberg, S. E., and Bokalders, K., Chapman & Hall, pp. 35–79.

Fraley, C. and Raftery, A. E. (1998), "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer Journal*, 41, 578–588.

— (2002), "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, 97, 611–631.

Frühwirth-Schnatter, S. and Kaufmann, S. (2008), "Model-based clustering of multiple time series," *Journal of Business & Economic Statistics*, 26, 78–89.

Fuentes, M. (2001), "A high frequency kriging approach for non-stationary environmental processes," *Environmetrics*, 12, 469–483.

Fuentes, M. and Raftery, A. E. (2005), "Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models," *Biometrics*, 61, 36–45.

Fuller, G. W. (2003), "Air quality in London 2003. Final report," `www.londonair.org.uk/london/reports/AirQualityInLondon2003.pdf`, last accessed date 25 June 2014.

Fuller, G. W., Carslw, D. C., and Lodge, H. W. (2002), "An empirical approach for the prediction of daily mean PM_{10} concentrations," *Atmospheric Environment*, 36, 1431–1441.

Fuller, G. W. and Green, D. (2006), "Evidence for increasing primary PM_{10} in London," *Atmospheric Environment*, 40, 6134–6145.

Gasparrini, A., Armstrong, B., and Kenward, M. G. (2010), "Distributed lag non-linear models," *Statistics in Medicine*, 29, 2224–2234.

Gelfand, A. E., Diggle, P., Fuentes, M., and Guttorp, P. (2010), *Handbook of Spatial Statistics*, Boca Raton, FL: Chapman & Hall/CRC.

Gelfand, A. E., Kim, H. K., Sirmans, C. F., and Banerjee, S. (2003), "Spatial modelling with spatially varying coefficient processes," *Journal of the American Statistical Association*, 98, 387–396.

Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001), "On the change of support problem for spatio-temporal data," *Biostatistics*, 2, 31–45.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, London: Chapman & Hall, 3rd ed.

Gelman, A. and Hill, J. (2007), *Data Analysis using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008), "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, 2, 1360–1383.

Ghahramani, Z. (2005), "Non-parametric Bayesian Methods. Uncertainty in Artificial Intelligence - Tutorial July 2005," `http://mlg.eng.cam.ac.uk/zoubin/talks/uai05tutorial-b.pdf`, last accessed date 27 August 2015.

Ghosal, S. (2010), "The Dirichlet Process, Related Priors and Posterior Asymptotics," in *Bayesian Nonparametrics*, eds. Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., Cambridge, UK: Cambridge University Press, pp. 35–79.

Ghosh, S. K., Bhave, P. V., Davis, J. M., and Lee, H. (2010), "Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models," *Journal of the American Statistical Association*, 105, 538–551.

Gilks, W. R. (1992), "Derivative-Free Adaptive Rejection Sampling for Gibbs Sampling," in *Bayesian Statistics*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford, UK: Oxford University Press, 4th ed., pp. 641–666.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Gilks, W. R. and Roberts, G. O. (1996), "Strategies for improving MCMC," in *Markov Chain Monte Carlo in Practice*, eds. Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., London: Chapman & Hall, pp. 89–114.

Goldman, G. T., Mulholland, J. A., Russell, A. G., Strickland, M. J., Klein, M., Waller, L. A., and Tolbert, P. E. (2011), "Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies," *Environmental Health*, 10, 61.

Gómez-Losada, A., Lozano-García, A., Pino-Mejías, R., and Contreras-González, J. (2014), "Finite mixture models to characterize and refine air quality monitoring networks," *Science of the Total Environment*, 485-486, 292–299.

Gotway, C. A. and Young, L. J. (2002), "Combining incompatible spatial data," *Journal of America Statistical Association*, 97, 632–648.

Green, D., Fuller, G., and Baker, T. (2009), "Development and validation of the volatile correction model for PM10 - An empirical method for adjusting TEOM measurements for their loss of volatile particulate matter," *Atmospheric Environment*, 43, 2132–2141.

Green, D., Fuller, G. W., and Barratt, B. (2001), "Evaluation of TEOM (TM) 'correction factors' for assessing the EU Stage 1 limit values for PM10," *Atmospheric Environment*, 35, 2589–2593.

Greenland, S. (1980), "The effect of misclassification in the presence of covariates," *American Journal of Epidemiology*, 112, 564–569.

Griffin, J. E. and Steel, M. F. J. (2006), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, 101, 179–194.

— (2011), "Stick-breaking autoregressive processes," *Journal of Econometrics*, 162, 383–396.

Griffiths, T. L. and Ghahramani, Z. (2006), "Infinite latent feature models and the Indian buffet process," in *Advances in Neural Information Processing Systems 18*, eds. Weiss, Y., Schölkopf, B., and Platt, J., MIT Press, pp. 475–482.

Gu, J., Pitz, M., Breitner, S., Birmili, W., von Klot, S., Schneider, A., Soentgen, J., Reller, A., Peters, A., and Cyrys, J. (2012), "Selection of key ambient particulate variables for epidemiological studies - Applying cluster and heatmap analyses as tools for data reduction," *Science of The Total Environment*, 435-436, 541–550.

Gulliver, J., Morris, C., Lee, K., Vienneau, D., Briggs, D., and Hansell, A. (2011), "Land use regression modeling to estimate historic (1962-1991) concentrations of black smoke and sulfur dioxide for Great Britain," *Environmental Science & Technology*, 45, 3526–3532.

Hamm, N. A. S., Finley, A. O., Schaap, M., and Stein, A. (2015), "A spatially varying coefficient model for mapping {PM10} air quality at the European scale," *Atmospheric Environment*, 102, 393–405.

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011), "Dirichlet process mixtures of generalized linear models," *Journal of Machine Learning Research*, 1, 1–33.

Harrison, D. (2006), "UK equivalence programme for monitoring of particulate matter," Tech. Rep. Ref: BV/AQ/AD202209/DH/2396, Department for Environment, Food & Rural Affairs, London.

Harrison, R. M. and Yin, J. (2000), "Particulate matter in the atmosphere: which particle properties are important for its effects on health?" *Science of The Total Environment*, 249, 85–101.

Hartigan, J. A. and Wong, M. A. (1979), "Algorithm AS 136: A K-means clustering algorithm," *Journal of the Royal Statistical Society: Series C*, 28, 100–108.

Hastie, D. I., Liverani, S., Azizi, L., Richardson, S., and Stücker, I. (2013), "A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer," *BMC Medical Research Methodology*, 13, 129.

Hastie, D. I., Liverani, S., and Richardson, S. (2014), "Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations," *Statistics and Computing*, 1–15.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York: Springer, 2nd ed.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman & Hall.

— (1993), "Varying-coefficient models," *Journal of the Royal Statistical Society: Series B*, 55, 757–796.

Hastings, W. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Heal, M. R. and Quincey, P. (2012), "The relationship between black carbon concentration and black smoke: a more general approach," *Atmospheric Environment*, 54, 538–544.

Heaton, M. J., Reese, C. S., and Christensen, W. F. (2010), "Incorporating time-dependent source profiles using the Dirichlet distribution in multivariate receptor models," *Technometrics*, 52, 67–79.

HEI (2002), *Understanding the health effects of components of the particulate matter mix: progress and next steps*, Boston, MA: Health Effects Institute.

HEI (2010), *Traffic-related air pollution: A critical review of the literature on emission, exposure, and health effects*, Boston, MA: Health Effect Institute, Special Report 17.

Henry, R. C. and Norris, G. A. (2002), *UNIMIC* 2.3 *User Guide*, U.S. EPA, Research Triangle Parc, NC.

Higdon, D. (1998), "A process-convolution approach to modelling temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190.

Hjort, N. L. (1990), "Nonparametric Bayes estimators based on beta processes in models for life history data," *Annals of Statistics*, 18, 1259–1294.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010), *Bayesian Nonparametrics*, Cambridge, UK: Cambridge University Press.

Hocking, R. R. and Pendelton, O. J. (1983), "The regression dilemma," *Communication in Statistics-Theory and Methods*, 12, 497–527.

Hoek, G., Beelen, R., de, H. K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D. (2008), "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric Environment*, 42, 7561–7578.

Hoerl, A. E. and Kennard, R. W. (1970), "Ridge regression: Applications to nonorthogonal problems," *Technometrics*, 12, 55–68.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian model averaging: A tutorial," *Statistical Science*, 14, 382–417.

Hong, Y. C., Leem, J. H., Ha, E. H., and Christiani, D. C. (1999), "PM_{10} exposure, gaseous pollutants, and daily mortality in Inchon, South Korea," *Environmental Health Perspectives*, 107, 873–878.

Hopke, P. K., Ito, K., Mar, T., Christensen, W. F., Eatough, D. J., Henry, R. C., Kim, E., Laden, F., Lall, R., Larson, T. V., Liu, H., Neas, L., Pinto,

J., Stolzel, M., Suh, H., Paatero, P., and Thurston, G. D. (2006), "PM source apportionment and health effects: 1. Intercomparison of source apportionment results," *Journal of Exposure Science and Environmental Epidemiology*, 16, 275–286.

Huang, Y., Dominici, F., and Bell, M. (2005), "Bayesian hierarchical distributed lag models for summer ozone exposure and cardio-respiratory mortality," *Environmetrics*, 16, 547–562.

Huerta, G., Sansó, B., and Stroud, J. R. (2004), "A spatio-temporal model for Mexico City ozone levels," *Journal of the Royal Statistical Society: Series C*, 53, 231–248.

IARC (2013), *IARC monographs on the evaluation of carcinogenic risks to humans. Volume 109. Outdoor air pollution*, Lyon, France: International Agency for Research on Cancer.

Ignaccolo, R., Ghigo, S., and Giovenali, E. (2008), "Analysis of air quality monitoring networks by functional clustering," *Environmetrics*, 19, 672–686.

Iorio, M. D., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA model for dependent random measures," *Journal of the American Statistical Association*, 99, 205–215.

Ito, K., Christensen, W. F., Eatough, D. J., Henry, R. C., Kim, E., Laden, F., Lall, R., Larson, T. V., Neas, L., Hopke, P. K., and Thurston, G. D. (2006), "PM source apportionment and health effects: 2. An investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC," *Journal of Exposure Science and Environmental Epidemiology*, 16, 300–310.

Jain, A. K. (2010), "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, 31, 651–666.

James, G., Witten, D., Hastie, T., and Tibshiran, R. (2013), *An Introduction to Statistical Learning: with Applications in R*, New York: Springer, 4th printing.

Janssen, N. A. H., Gerlofs-Nijland, M. E., Lanki, T., Salonen, R. O., Cassee, F., Hoek, G., Paul Fische and, B. B., and Krzyzanowski, M. (2012), "Health effects of black carbon," Tech. rep., WHO Regional Office for Europe, Copenhagen, Denmark.

Johnson, H., Kovats, S., McGregor, G., Stedman, J., Gibbs, M., and Walton, H. (2005), "The impact of the 2003 heat wave on daily mortality in England and Wales and the use of rapid weekly mortality estimates," *Euro Surveilance*, 10.

Johnson, S. C. (1967), "Hierarchical clustering schemes," *Psychometrika*, 2, 241–254.

Jones, A. M., Harrison, R. M., Barratt, B., and Fuller, G. W. (2012), "A large reduction in airborne particle number concentrations at the time of the introduction of "sulphur free" diesel and the London Low Emission Zone," *Atmospheric Environment*, 50, 129–138.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21, 93–105.

Kamruzzaman, M. D. and Imon, A. H. M. R. (2002), "High leverage point: another source of multicollinearity," *Pakistan Journal of Statistics*, 18, 435–448.

Kan, H. D. and Chen, B. H. (2004), "Statistical distributions of ambient air pollutants in Shanghai, China," *Biomedical and Environmental Sciences*, 17, 366–372.

Kang, C. and Ghosal, S. (2009), "Clusterwise Regression using Dirichlet Mixtures," in *In Advances in Multivariate Statistical Methods*, ed. Sengupta, A., World Scientific Publishing Company, pp. 305–325.

Kaufman, L. and Rousseeuw, P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: John Wiley & Sons.

Kavitha, T. V. and Punithavalli, M. (2010), "Clustering time series data stream - A literature survey," *International journal of Computer Science and Information Security*, 8, 289–293.

Kelly, F. J. and Fussell, J. C. (2012), "Size, source and chemical composition as determinants of toxicity attributable to ambient particulate matter," *Atmospheric Environment*, 60, 504–526.

Kim, E., Hopke, P. K., and Edgerton, E. S. (2003), "Source identification of Atlanta aerosol by positive matrix factorization," *Journal of the Air & Waste Management Association*, 53, 731–739.

Kim, K.-H., Kabir, E., and Kabir, S. (2015), "A review on the human health impact of airborne particulate matter," *Environment International*, 74, 136–143.

Koop, G. and Tole, L. (2004), "Measuring the health effects of air pollution: to what extent can we really say that people are dying from bad air?" *Journal of Environmental Economics and Management*, 47, 30–54.

Krzyzanowski, M., Kuna-Dibbert, B., and Schneider, J. (2005), "Health effects of transport-related air pollution," Tech. rep., WHO Regional Regional Office for Europe, Copenhagen, Denmark.

Laurent, O., Pedrono, G., Segala, C., Filleul, L., Havard, S., Deguen, S., Schillinger, C., RiviÃ¨re, E., and Bard, D. (2008), "Air pollution, asthma attacks, and socioeconomic deprivation: a small-area case-crossover study," *American Journal of Epidemiology*, 168, 58–65.

Le, N. D. and Zidek, J. V. (2006), *Statistical Analysis of Environmental Space-Time Processes*, New York: Springer.

Lee, D., Ferguson, C., and Scott, E. M. (2011), "Constructing representative air quality indicators with measures of uncertainty," *Journal of the Royal Statistical Society: Series A*, 174, 109–126.

Lee, D. and Shaddick, G. (2007), "Time-varying coefficient models for the analysis of air pollution and health outcome data," *Biometrics*, 63, 1253–1261.

— (2008), "Modelling the effects of air pollution on health using Bayesian dynamic generalised linear models," *Environmetrics*, 19, 785–804.

— (2010), "Spatial modeling of air pollution in studies of its short-term health effects," *Biometrics*, 66, 1238–1246.

Lenschow, P., Abraham, H. J., Kutzner, K., Lutz, M., Preuß, J. D., and Reichenbächer, W. (2001), "Some ideas about the sources of PM10," *Atmospheric Environment*, 35, S23–S33.

Liao, T. W. (2005), "Clustering of time series data - A survey," *Pattern Recognition*, 38, 1857–1874.

Lijoi, A. and Prünster, I. (2010), "Models beyond the Dirichlet process," in *Bayesian Nonparametrics*, eds. Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., Cambridge, UK: Cambridge University Press, pp. 35–79.

Lindgren, F. and Rue, H. (2008), "On the second-order random walk model for irregular locations," *Scandinavian Journal of Statistics*, 35, 691–700.

Lingwall, J. W., Christensen, W. F., and Reese, C. S. (2008), "Dirichlet based Bayesian multivariate receptor modeling," *Environmetrics*, 19, 618–629.

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., and Richardson, S. (2015), "PReMiuM: an R package for profile regression mixture models using Dirichlet processes," *Journal for Statistical Software*, 64, 1–30.

Logan, W. P. D. (1953), "Mortality in the London fog incident, 1952," *The Lancet*, 261, 336–338.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., Straif, K., and on

behalf of the International Agency for Research on Cancer Monograph Working Group IARC (2013), "The carcinogenicity of outdoor air pollution," *The Lancet Oncology*, 14, 1262–1263.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility," *Statistics and Computing*, 10, 325–337.

MacLehose, R. F., Dunson, D., Herring, A. H., and Hoppin, J. A. (2007), "Bayesian methods for highly correlated exposure data," *Epidemiology*, 18, 199–207.

Mar, T. F., Ito, K., Koenig, J. Q., Larson, T. V., Eatough, D. J., Henry, R. C., Kim, E., Laden, F., Lall, R., Neas, L., Stolzel, M., Paatero, P., Hopke, P. K., and Thurston, G. D. (2006), "PM source apportionment and health effects. 3. Investigation of inter-method variations in associations between estimated source contributions of $PM_{2.5}$ and daily mortality in Phoenix, AZ," *Journal of Exposure Science and Environmental Epidemiology*, 16, 311–320.

Marquardt, D. W. (1970), "Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation," *Technometrics*, 12, 591–612.

Martin, M. A. and Roberts, S. (2006), "Boostrap model averaging time series studies of particulate matter air pollution and mortality," *Journal of Exposure Science and Environmental Epidemiology*, 16, 242–250.

Marx, B. D. (2010), "P-spline Varying Coefficient Models for Complex Data," in *Statistical Modelling and Regression Structures*, eds. Kneib, T. and Tutz, G., Berlin: Springer-Verlag, pp. 19–43.

Marx, B. D. and Eilers, P. H. C. (1998), "Direct generalized additive modeling with penalized likelihood," *Computational Statistics and Data Analysis*, 28, 193–209.

Matérn, B. (1986), *Spatial Variation*, Berlin: Springer-Verlag, 2nd ed.

Matheron, G. (1962), *Traité de Géostatistique Appliquée*, Editions Technip.

Matyasovszky, I., Makra, L., Bálint, B., Guba, Z., and Sümeghy, Z. (2011), "Multivariate analysis of respiratory problems and their connection with meteorological parameters and the main biological and chemical air pollutants," *Atmospheric Environment*, 45, 4152–4159.

Mauderly, J. L., Burnett, R. T., Castillejos, M., Ozkaynak, H., Samet, J. M., Stieb, D. M., Vedal, S., and Wyzga, R. E. (2010), "Is the air pollution health research community prepared to support a multipollutant air quality management framework?" *Inhalation Toxicology*, 22, 1–19.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall, 2nd ed.

McHugh, C. A., Carruthers, D. J., and Edmunds, H. A. (1997), "ADMS and ADMS-Urban," *International Journal of Environment and Pollution*, 8, 438–440.

McLachlan, G. J. and Baek, J. (2010), "Clustering of high-dimensional data via finite mixture models," in *Advances in Data Analysis, Data Handling and Business Intelligence*, eds. Fink, A., Lausen, B., Seidel, W., and Ultsch, A., Springer-Verlag, pp. 33–44.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, John Wiley & Sons, Inc.

McMillan, N., Holland, D. M., Morara, M., and Feng, J. (2010), "Combining numerical model output and particulate data using Bayesian space-time modeling," *Environmetrics*, 21, 48–65.

McMurry, P. H., Shepherd, M. F., and Vickery, J. S. (2004), *Particulate Matter Science for Policy Makers: A NARSTO Assessment*, Cambridge University Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1092.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010), "Bayesian profile regression with an application to the National Survey of Children's Health," *Biostatistics*, 11, 484–498.

Molitor, J., Su, J. G., Molitor, N. T., Rubio, V. G., Richardson, S., Hastie, D., Morello-Frosch, R., and Jerrett, M. (2011), "Identifying vulnerable populations through an examination of the association between multipollutant profiles and poverty," *Environmental Science & Technology*, 45, 7754–7760.

Moller, S. F., Frese, J. V., and Bro, R. (2005), "Robust methods for multivariate data analysis," *Journal of Chemometrics*, 19, 549–563.

Monn, C. (2001), "Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone," *Atmospheric Environment*, 35, 1–32.

Montgomery, D. C., Peck, P. P., and Vining, G. G. (2001), *Introduction to Linear Regression Analysis*, New York: John Wiley & Sons, Inc, 3rd ed.

Morawska, L., Afshari, A., Bae, G. N., Buonanno, G., Chao, C. Y. H., Hänninen, O., Hofmann, W., Isaxon, C., Jayaratne, E. R., Pasanen, P., Salthammer, T., Waring, M., and Wierzbicka, A. (2013), "Indoor aerosols: from personal exposure to risk assessment," *Indoor Air*, 23, 462–487.

Morlini, I. (2006), "On multicollinearity and concurvity in some nonlinear multivariate models," *Statistical Methods and Applications*, 15, 3–26.

Mortimer, K., Neugebauer, R., Lurmann, F., Alcorn, S., Balmes, J., and Tager, I. (2008), "The effect of prenatal and lifetime exposure to air pollution on the pulmonary function of asthmatic children," *Epidemiology*, 19, 550–557.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Müller, P. and Mitra, R. (2013), "Bayesian nonparametric inference - Why and how," *Bayesian Analysis*, 8, 269–302.

Müller, P. and Quintana, F. A. (2010), "Random partition models with regression on the covariates," *Journal of Statistical Planning and Inference*, 140, 2801–2808.

Müller, P., Quintana, F. A., and Rosner, G. (2004), "A method for combining inference across related nonparametric Bayesian models," *Journal of the Royal Statistical Society: Series B*, 66, 735–749.

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*, Massachusetts Institute of Technology.

National Research Council (2004), *Research priorities for airborne particulate matter: IV. Continuing research progress*, Washington, DC: The National Academies Press.

— (2007), *Models in environmental regulatory decision making*, Washington, DC: The National Academies Press.

Neal, R. M. (1992), "Bayesian mixture modeling," in *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, Seattle, pp. 197–211.

— (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

— (2003), "Slice sampling," *Annals of Statistics*, 31, 705–767.

Nieto-Barajas, L. E. and Contreras-Cristán, A. (2014), "A Bayesian nonparametric approach for time series clustering," *Bayesian Analysis*, 9, 147–170.

Oakes, M., Baxter, L., and Long, T. C. (2014), "Evaluating the application of multipollutant exposure metrics in air pollution health studies," *Environment International*, 69, 90–99.

O'Hara, R. B. and Sillanpää, M. J. (2009), "A review of Bayesian variable selection methods: what, how and which," *Bayesian Analysis*, 4, 85–117.

Orbanz, P. and Teh, Y. (2010), "Bayesian Nonparametric Models," in *Encyclopedia of Machine Learning*, Springer.

Ostro, O., Roth, L., Malig, B., and Marty, M. (2009), "The effects of fine particle components on respiratory hospital admissions in children," *Environmental Health Perspectives*, 117, 475–180.

Ott, W. R. (1990), "A physical explanation of the lognormality of pollutant concentrations," *Journal of the Air Waste Management Association*, 40, 1378–1383.

Özkaynak, H., Baxter, L. K., Dionisio, K. L., and Burke, J. (2013), "Air pollution exposure prediction approaches used in air pollution epidemiology studies," *Journal of Exposure Science and Environmental Epidemiology*, 23, 566–572.

Paatero, P. and Tapper, U. (1994), "Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, 5, 111–126.

Papaspiliopoulos, O. and Roberts, G. O. (2008), "Retrospective Markov chain Monte Carlo for Dirichlet process hierarchical models," *Biometrika*, 95, 169–186.

Papathomas, M., Molitor, J., Hoggart, C., Hastie, D., and Richardson, S. (2012), "Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene x gene patterns," *Genetic Epidemiology*, 6, 663–674.

Papathomas, M., Molitor, J., Richardson, S., Riboli, E., and Vineis, P. (2011), "Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers," *Environmental Health Perspectives*, 119, 84–91.

Park, B. U., Mammen, E., Lee, Y. K., and Lee, E. R. (2015), "Varying coefficient regression models: a review and new developments," *International Statistical Review*, 83], pages = 36–64,.

Park, E. S., Guttorp, P., and Henry, R. C. (2001), "Multivariate receptor modeling for temporally correlated data by using MCMC," *Journal of the American Statistical Association*, 96, 1171–1183.

Park, E. S., Oh, M.-S., and Guttorp, P. (2002), "Multivariate receptor models and model uncertainty," *Chemometrics and Intelligent Laboratory Systems*, 60, 49–67.

Peng, R. D. (2008), "A method for visualizing multivariate time series," *Journal of Statistical Software*, 25, 1–17.

Peng, R. D. and Bell, M. L. (2010), "Spatial misalignment in time series studies of air pollution and health data," *Biostatistics*, 11, 720–740.

Peng, R. D., Bell, M. L., Geyh, A. S., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2009), "Emergency admission for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution," *Environmental Health Perspectives*, 117, 957–963.

Peng, R. D. and Dominici, F. (2008), *Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health*, New York: Springer.

Peng, R. D., Dominici, F., and Louis, T. A. (2006), "Model choice in time series studies of air pollution and mortality," *Journal of Royal Statistical Society: Series A*, 169, 179–203.

Peng, R. D., Dominici, F., Pastor-Barriuso, R., Zeger, S. L., and Samet, J. M. (2005), "Seasonal analyses of air pollution and mortality in 100 US cities," *American Journal of Epidemilogy*, 161, 585–594.

Pirani, M., Best, N., Blangiardo, M., Liverani, S., Atkinson, R. W., and Fuller, G. W. (2015), "Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles," *Environment International*, 79, 56–64.

Pirani, M., Gulliver, J., Fuller, G. W., and Blangiardo, M. (2014), "Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas," *Journal of Exposure Science and Environmental Epidemiology*, 24, 319–327.

Pitard, A. and Viel, J. F. (1997), "Some methods to address collinearity among pollutants in epidemiological time series," *Statistics in Medicine*, 16, 527–544.

Pitman, J. (1995), "Exchangeable and partially exchangeable random partitions," *Probability Theory and Related Fields*, 102, 145–158.

— (2006), "Combinatorial stochastic processes," *U.C. Berkeley, Department of Statistics*, 621.

Pollice, A. (2011), "Recent statistical issues in multivariate receptor models," *Environmetrics*, 22, 35–41.

Powell, H. and Lee, D. (2014), "Modelling spatial variability in concentrations of single pollutants and composite air quality indicators in health effects studies," *Journal of the Royal Statistical Society: Series A*, 177, 607–623.

Prado, R. and West, M. (2010), *Time Series: Modeling, Computation, and Inference*, Boca Raton, FL: Chapman & Hall/CRC.

Pun, V. C., Yu, I. T., Qiu, H., Ho, K. F., Sun, Z., Louie, P. K., Wong, T. W., and Tian, L. (2014), "Short-term associations of cause-specific emergency hospital-

izations and particulate matter chemical components in Hong Kong," *American Journal of Epidemiology*, 179, 1086–1095.

Putaud, J.-P., Van Dingenen, R., Alastuey, A., Bauer, H., Birmili, W., Cyrys, J., Flentje, H., Fuzzi, S., Gehrig, R., Hansson, H. C., Harrison, R. M., Herrmann, H., Hitzenberger, R., Hüglin, C., Jones, A. M., Kasper-Giebl, A., Kiss, G., Kousa, A., Kuhlbusch, T. A. J., Löschau, G., Maenhaut, W., Molnar, A., Moreno, T., Pekkanen, J., Perrino, C., Pitz, M., Puxbaum, H., Querol, X., Rodriguez, S., Salma, I., Schwarz, J., Smolik, J., Schneider, J., Spindler, G., ten Brink, H., Tursic, J., Viana, M., Wiedensohler, A., and Raes, F. (2010), "A European aerosol phenomenology - 3: Physical and chemical characteristics of particulate matter from 60 rural, urban, and kerbside sites across Europe," *Atmospheric Environment*, 44, 1308–1320.

Ramsay, T. O., Burnett, R. T., and Krewski, D. (2003), "The effect of concurvity in generalized additive models linking mortality to ambient particulate matter," *Epidemiology*, 14, 18–23.

Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66, 846–850.

Rasmussen, C. E. and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, Cambridge, MA, USA: MIT Press.

Reich, B. J., Fuentes, M., and Burke, J. (2009), "Analysis of the effects of ultrafine particulate matter while accounting for human exposure," *Environmetrics*, 20, 131–146.

Reiss, R., Anderson, E. L., Cross, C. E., Hidy, G., Hoel, D., McClellan, R., and Moolgavkar, S. (2007), "Evidence of health impacts of sulfate-and nitrate-containing particles in ambient air," *Inhalation Toxicology*, 19, 419–449.

Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with

an unknown number of components (with discussion)," *Journal of the Royal Statistical Society: Series B*, 59, 731–792.

Roberts, S. (2006), "A new model for investigating the mortality effects of multiple air pollutants in air pollution mortality time-series studies," *Journal of Toxicology and Environmental Health, Part A*, 69, 417–435.

Roberts, S. and Martin, M. A. (2005), "Shrinkage-based regression approaches for estimating the adverse health effects of multiple air pollutants," *Atmospheric Environment*, 39, 6223–6230.

— (2006a), "Investigating the mixture of air pollutants associated with adverse health outcomes," *Atmospheric Environment*, 40, 984–991.

— (2006b), "Using supervised principal components analysis to assess multiple pollutant effects," *Environmental Health Perspective*, 114, 1877–1882.

— (2010), "Bootstrap-after-bootstrap model averaging for reducing model uncertainty in model selection for air pollution mortality studies," *Environmental Health Perspectives*, 118, 131–136.

Rodríguez, A. and Dunson, D. B. (2011), "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Analysis*, 6, 145–177.

Romieu, I., Gouveia, N., Cifuentes, L. A., de Leon, A. P., Junger, W., Vera, J., Strappa, V., Hurtado-Díaz, M., Miranda-Soberanis, V., Rojas-Bracho, L., Carbajal-Arroyo, L., Tzintzun-Cervantes, G., and HEI Health Review Committee (2012), *Multicity study of air pollution and mortality in Latin America (the ESCALA Study)*, Boston, MA: Health Effects Institute, Research Report 171.

Roscoe, B. A., Hopke, P. K., Dattner, S. L., and Jenks, J. M. (1982), "The use of principal component factor analysis to interpret particulate compositional data sets," *Journal of Air Pollution Control Association*, 32, 637–642.

Royle, J. A. and Dorazio, R. M. (2008), *Hierarchical Modeling and Inference in Ecology*, San Diego: Academic Press.

Ruppert, D. (2002), "Selecting the number of knots for penalized splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D. and Carroll, R. J. (2000), "Spatially-adaptive penalties for spline fitting," *Australian and New Zealand Journal of Statistics*, 42, 205–224.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2006), "Spatio-temporal modeling of fine particulate matter," *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 61–86.

— (2007), "High resolution space-time ozone modeling for assessing trends," *Journal of the American Statistical Association*, 102, 1221–1234.

— (2010), "Fusing point and areal level space-time data with application to wet deposition," *Journal of the Royal Statistical Society: Series C*, 59, 77–103.

Sahu, S. K., Yip, S., and Holland, D. M. (2009), "Improved space-time forecasting of next day ozone concentrations in the eastern US," *Atmospheric Environment*, 43, 494–501.

Samet, J. M., Dominici, F., Zeger, S. L., Schwartz, J., and Dockery, D. W. (2000a), *The National Morbidity, Mortality, and Air Pollution Study. Part I: Methods and methodologic issues*, Cambridge, MA: Health Effects Institute, Research Report 94.

Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D. W., and Schwartz, J. (2000b), *The National Morbidity, Mortality, and Air Pollution Study. Part II: Morbidity and mortality from air pollution in the United States*, Cambridge, MA: Health Effects Institute, Research Report 94.

Samoli, E., Zanobetti, A., Schwartz, J., Atkinson, R., LeTertre, A., Schindler, C., Perez, L., Cadum, E., Pekkanen, J., Pald, A., Touloumi, G., and Katsouyanni, K. (2009), "The temporal pattern of mortality reponses to ambient ozone in the APHEA project," *Journal of Epidemiology and Community Health*, 63, 960–966.

Sarnat, J. A., Marmur, A., Klein, M., Kim, E., Russell, A. G., Sarnat, S. E., Mulholland, J. A., Hopke, P. K., and Tolbert, P. E. (2008), "Fine particle sources and cardiorespiratory morbidity: an application of chemical mass balance and factor analytical source-apportionment methods," *Environmental Health Perspectives*, 116, 459–466.

Sarnat, J. A., Wilson, W. E., Strand, M., Brook, J., Wyzga, R., and Lumley, T. (2007), "Panel discussion review: session 1 - Exposure assessment and related errors in air pollution epidemiologic studies," *Journal of Exposure Science and Environmental Epidemiology*, 17, S75–S82.

Sarnat, S. E., Klein, M., Sarnat, J. A., Flanders, W. D., Waller, L. A., Mulholland, J. A., Russell, A. G., and Tolbert, P. E. (2010), "An examination of exposure measurement error from air pollutant spatial variability in time-series studies," *Journal of Exposure Science and Environmental Epidemiology*, 20, 135–146.

Schaap, M., van Loon, M., ten Brink, H. M., Dentener, F. J., and Builtjes, P. J. H. (2004), "Secondary inorganic aerosol simulations for Europe with special attention to nitrate," *Atmospheric Chemistry and Physics*, 4, 857–874.

Schabenberge, O. and Gotway, C. A. (2004), *Statistical Methods for Spatial Data Analysis*, Chapman and Hall/CRC.

Schauer, J. J., Lough, G. C., Shafer, M. M., Christensen, W. F., Arndt, M. F., DeMinter, J. T., and Park, J.-S. (2006), *Characterization of metals emitted from motor vehicles*, Boston, MA: Health Effect Institute, Research Report 133.

Schauer, J. J., Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. (1996), "Source apportionment of airborne particulate matter using organic compounds as tracers," *Atmospheric Environment*, 30, 3837–3855.

Schwartz, J. and Marcus, A. (1990), "Mortality and air pollution in London: a time series analysis," *American Journal of Epidemiology*, 131, 185–194.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Scott, J. A. (1953), "Fog and deaths in London, December 1952," *Public Health Reports*, 68, 474–479.

Seinfeld, J. H. and Pandis, S. N. (2006), *Atmospheric Chemistry and Physics: from Air Pollution to Climate Change*, New York: John Wiley & Sons, Inc, 2nd ed.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Shaddick, G., Lee, D., and Wakefield, J. (2013), "Ecological bias in studies of the short-term effects of air pollution on health," *International Journal of Applied Earth Observation and Geoinformation*, 22, 65–74.

Shaddick, G. and Wakefield, J. (2002), "Modelling daily multivariate pollutant data at multiple sites," *Journal of the Royal Statistical Society: Series C*, 51, 351–372.

Shahbaba, B. and Neal, R. (2009), "Nonlinear models using Dirichlet process mixtures," *Journal of Machine Learning Research*, 10, 1829–1850.

Sheppard, L., Burnett, R. T., Szpiro, A. A., Kim, S.-Y., Jerrett, M., Pope III, C. A., and Brunekreef, B. (2012), "Confounding and exposure measurement

error in air pollution epidemiology," *Air Quality, Atmosphere, & Health*, 5, 203–216.

Shieh, Y.-Y. and Fouladi, R. T. (2003), "The effect of multicollinearity on multilevel modeling parameter estimates and standard errors," *Educational and Psychological Measurement*, 63, 951.

Shorack, G. R. and Wellner, J. (1986), *Empirical Processes with Applications in Statistics*, New York: Wiley.

Shumway, R. H. and Stoffer, D. S. (2011), *Time Series Analysis and Its Applications. With R Examples*, Springer, 3rd ed.

Sinisi, S. E. and van der Laan, M. J. (2004), "Deletion/substitution/addition algorithm in learning with applications in genomics," *Statistical Applications in Genetics and Molecular Biology*, 3.

Sjöström, M., Wold, S., Lindberg, W., Persson, J.-A., and Martens, H. (1983), "A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables," *Analytica Chimica Acta*, 150, 61–70.

Solberg, S., Coddeville, P., Forster, C., Hov, Ø., Orsolini, Y., and Uhse, K. (2005), "European surface ozone in the extreme summer 2003," *Atmospheric Chemistry and Physics*, 5, 9003–9038.

Stanek, L. W., Brown, J. S., Stanek, J., Gift, J., and Costa, D. L. (2011a), "Air pollution toxicology - A brief review of the role of the science in shaping the current understanding of air pollution health risks," *Toxicological Sciences*, 120, S8–S27.

Stanek, L. W., Sacks, J. D., Dutton, S. J., and Dubois, J.-J. B. (2011b), "Attributing health effects to apportioned components and sources of particulate matter: an evaluation of collective results," *Atmospheric Environment*, 45, 5655–5663.

Sun, Z., Tao, Y., Li, S., Ferguson, K. K., Meeker, J. D., Park, S. K., Batterman, S. A., and Mukherjee, B. (2013), "Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons," *Environmental Health*, 12, 85.

Taddy, M. A. and Kottas, A. (2009), "Markov switching Dirichlet process mixture regression," *Bayesian Analysis*, 4, 793–816.

Tang, R., Blangiardo, M., and Gulliver, J. (2013), "Using building heights and street configuration to enhance intraurban PM_{10}, NO_X, and NO_2 land use regression models," *Environmental Science & Technology*, 47, 11643–11650.

Taylor, K. (2001), "Summarizing multiple aspects of model performance in a single diagram," *Journal of Geophysical Research: Atmospheres*, 106, 7183–7192.

Teh, Y. W. (2010), "Dirichlet Processes," in *Encyclopedia of Machine Learning*, Springer.

Tenías, J. M., Ballester, F., and Rivera, M. L. (1998), "Association between hospital emergency visits for asthma and air pollution in Valencia, Spain," *Occupational and Environmental Medicine*, 55, 36–52.

Thomas, D. C., Jerrett, M., Kuenzli, N., Louis, T. A., Dominici, F., Zeger, S., Schwarz, J., Burnett, R. T., Krewski, D., and Bates, D. (2007a), "Bayesian model averaging in time-series studies of air pollution and mortality," *Journal of Toxicology and Environmental Health, Part A*, 18, 186–190.

Thomas, D. C., Witte, J. S., and Greenland, S. (2007b), "Dissecting effects of complex mixtures: who's afreid of informative priors?" *Epidemiology*, 18, 186–190.

Thurston, G. D., Ito, K., Mar, T., Christensen, W. F., Eatough, D. J., Henry, R. C., Kim, E., Laden, F., Lall, R., Larson, T. V., Liu, H., Neas, L., Pinto, J., Stolzel, M., Suh, H., and Hopke, P. K. (2005), "Workgroup report: workshop

on source apportionment of particulate matter health effects-intercomparison of results and implications," *Environmental Health Perspectives*, 113, 1768–1774.

Thurston, G. D. and Spengler, J. D. (1985), "A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston," *Atmospheric Environment*, 19, 9–25.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

Touloumi, G., Samoli, E., and Katsouyanni, K. (1996), "Daily mortality and "winter type" air pollution in Athens, Greece - A time series analysis within the APHEA project," *Journal of Epidemiology and Community Health*, 50, s47–s51.

Touloumi, G., Samoli, E., Pipikou, M., Le Tertre, A., Atkinson, R., and Katsouyanni, K. (2006), "Seasonal confounding in air pollution and health time-series studies: effect on air pollution effect estimates," *Statistics in Medicine*, 25, 4164–4178.

Turner, P. (2000), *Guide to Scientific Computing*, CRC Press, 2nd ed.

U.S. EPA (1996), *Air quality criteria for particulate matter*, Washington, DC: U.S. Environmental Protection Agency, volume III.

— (2008), *The multi-pollutant report: technical concepts and examples*, Washington, DC: U.S. Environmental Protection Agency.

— (2012), *Provisional assessment of recent studies on health effects of particulate matter exposure*, Washington, DC: U.S. Environmental Protection Agency, ePA/600/R-12/056.

Vedal, S. and Kaufman, J. D. (2011), "What does multi-pollutant air pollution research mean?" *American Journal of Respiratory and Critica Care Medicine*, 183, 4–6.

Viana, M., Kuhlbusch, T., Querol, X., Alastuey, A., Harrison, R., Hopke, P., Winiwarter, W., Vallius, M., Szidat, S., Prevot, A., Hueglin, C., Bloemen, H., Wahlin, P., Vecchi, R., Miranda, A., Kasper-Giebl, A., Maenhaut, W., and Hitzenberger, R. (2008), "Source apportionment of particulate matter in Europe: a review of methods and results," *Journal of Aerosol Science*, 39, 827–849.

Wade, S., Dunson, D. B., Petrone, S., and Trippa, L. (2014), "Improving prediction from Dirichlet process mixtures via enrichment," *Journal of Machine Learning Research*, 15, 1041–1071.

Wahba, G. (1990), *Spline models for observational data*, Philadelphia: Society for Industrial and Applied Mathematics.

Walker, S. G. (2007), "Sampling the Dirichlet mixture model with slices," *Communications in Statistics - Simulation and Computation*, 36, 45–54.

Wang, C., Parmigiani, G., and Dominici, F. (2012), "Bayesian effect estimation accounting for adjustment uncertainty," *Biometrics*, 68, 661–671.

Watson, J. G. and Schoen, R. (1984), "The effective variance weighting for least squares calculations applied to the mass balance receptor model," *Atmospheric Environment*, 18, 1347–1355.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, New York: Springer-Verlag, 2nd ed.

Whitby, K. T. (1978), "The physical characteristics of sulfur aerosols," *Atmospheric Environment*, 12, 135–159.

Whitby, K. T., Liu, B. Y. H., Husar, R. B., and Barsic, N. J. (1972), "The Minnesota aerosol-analyzing system used in the Los Angeles smog project," *Journal of Colloid and Interface Science*, 39, 136–164.

WHO (2014), *Burden of Disease from Ambient Air Pollution for 2012*, Geneva, Switzerland.

WHO/Europe (1999), *Monitoring Ambient Air Quality for Health Impact Assessment*, Copenhagen, Denmark: WHO Regional Office for Europe.

— (2000), *Air Quality Guidelines*, Copenhagen, Denmark: WHO Regional Office for Europe, 2nd ed.

— (2004), *Health aspects of air pollution. Results from the WHO project "Sistematic review of health aspects of air pollution in Europe"*, Copenhagen, Denmark: WHO Regional Office for Europe.

— (2006), *Health Risks of Particulate Matter from Long-range Transboundary Air Pollution*, Copenhagen, Denmark: WHO Regional Office for Europe.

— (2007), *Health Relevance of Particulate Matter from Various Sources*, Copenhagen, Denmark: WHO Regional Office for Europe.

— (2013), *Review of Evidence on Health Aspects of Air Pollution - REVIHAAP Project*, Copenhagen, Denmark: WHO Regional Office for Europe.

Wickle, C. K. (2003), "Hierarchical models in environmental science," *International Statistical Review*, 71, 181–199.

Wikle, C. K. and Berliner, L. M. (2005), "Combining information across spatial scales," *Technometric*, 47, 80–91.

— (2007), "A Bayesian tutorial for data assimilation," *Physica D-nonlinear Phenomena*, 230, 1–16.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001), "Spatiotemporal hierarchical Bayesian modeling: tropical ocean surface winds," *Journal of the American Statistical Association*, 96, 382–397.

Zanobetti, A., Austin, E., Coull, B. A., Schwartz, J., and Koutrakis, P. (2014), "Health effects of multi-pollutant profiles," *Environment International*, 71, 13–19.

Zanobetti, A., Franklin, M., Koutrakis, P., and Schwartz, J. (2009), "Fine particulate air pollution and its components in association with cause-specific emergency admissions," *Environmental Health*, 8, 58.

Zanobetti, A., Schwartz, J., Samoli, E., Gryparis, A., Touloumi, G., Atkinson, R., Le Tertre, A., Bobros, J., Celko, M., Goren, A., Forsberg, B., Michelozzi, P., Rabczenko, D., Aranguez Ruiz, E., and Katsouyanni, K. (2002), "The temporal pattern of mortality responses to air pollution: a multicity assessment of mortality displacement," *Epidemiology*, 13, 87–93.

Zanobetti, A., Wand, M. P., Schwartz, J., and Ryan, L. M. (2000), "Generalized additive distributed lag models: quantifying mortality displacement," *Biostatistics*, 1, 279–292.

Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., and Cohen, A. (2000), "Exposure measurement error in time-series studies of air pollution: concepts and consequences," *Environmental Health Perspectives*, 108, 419–426.

Zeka, A. and Schwartz, J. (2004), "Estimating the independent effects of multiple pollutants in the presence of measurement error: an application of a measurement-error-resistant technique," *Environmental Health Perspectives*, 112, 1686–1690.

Zhao, Q., Liang, Z., Tao, S., Zhu, J., and Du, Y. (2011), "Effects of air pollution on neonatal prematurity in Guangzhou of China: a time-series study," *Environmental Health*, 10, 2.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B*, 67, 301–320.